UNIVERSITAT DE BARCELONA

Biological Applications of Discrete Molecular Dynamics

Pedro Sfriso

UNIVERSITAT DE BARCELONA

INSTITUTE FOR RESEARCH IN BIOMEDICINE

Doctoral Programme in Biomedicine

University of Barcelona

2015

Biological Applications of Discrete Molecular Dynamics

Research leading to this Thesis was performed in the Molecular Modeling and Bioinformatics Lab at Institute for Research in Biomedicine IRB Barcelona, in the Joint BSC--IRB Research Program in Computational Biology.

Pedro Sfriso

PhD Candidate

Prof. Dr. Modesto Orozco

PhD Advisor

Acknowledgements

Cuando por tercera, buscando bibliografía, encuentras el trabajo previo de tu supervisor te das cuenta que has aprendido de él la forma de resolver problemas en ciencia. Sólo puedo agradecerle repetidamente a Modesto Orozco su paciencia en este camino. En la misma línea, Patrick Aloy y Josep Lluís Gelpí me animaron a explorar otros mundos con su incalculable experiencia.

Agustí me inció en el mundo de la simulaciones. Adam me mostró como se cambia la escala de un problema. Marcos Villarreal nos empujó a todos con su astucia. Michela e Ivan me auparon. Laura me mostró problemas reales y Miquel a volver a empezar, una y otra vez hasta que no te salga mejor. Más de la mitad de lo conseguido en esta Tesis es vuestro mérito.

I would like to thank Prof. Ron Elber for hosting me in his Lab. His particular point of view felt like a start over for me.

Gracias.

Preface

I started this Thesis with the idea of finding original applications of discrete Molecular Dynamics simulation method, but I have got stuck in the first biological process I attempted to simulate: conformational changes of proteins. The reader is alerted that some statements in this Thesis do not apply exclusively to simulations of such processes, but can be of larger scope. For simplicity, I left aside other macromolecules, other cellular processes and other methods along the text. I apologize for that.

I have to disclose that results are presented as if they were obtained following a designed path. Obviously, this was not the case, complicated detail, tedious problems and weaker points, were there, but are reserved for corridor talks.

This Thesis is about methodological developments; therefore the main plot is traced with methods limitations, advantages and methods tricky sides. I hope that I manage to let the reader share my enthusiasm about basic theory more than bore her/him with technical details.

Table of Contents

PhD Advisor Report

This section is included following mandatory guidelines.

Publications

Sfriso, P; Emperador, A; Orellana, L; Hospital, A; Gelpi, JL and Orozco, M. (2012) Finding Conformational Transition Pathways from Discrete Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* 8:4707–4718.

> P. Sfriso was the main developer of the project and contributed to the writing of the paper.

Sfriso P, Hospital A, Emperador A, Orozco M (2013) Exploration of conformational transition pathways from coarse-grained simulations. *Bioinformatics* 29:1980–1986.

> P. Sfriso was the main developer of the project and contributed to the writing of the paper.

Emperador, A; Solernou, A; Sfriso, P; Pons, C; Gelpi, JL; Fernandez-Recio, J and Orozco, M. (2013) Efficient Relaxation of Protein–Protein Interfaces by Discrete Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* 9:1222–1229.

> P. Sfriso designed the scoring function and energy potentials, performed some simulations and contributed to the writing of the paper.

Emperador A, Sfriso P, Villarreal MA, Gelpí JL, Orozco M (2015) PACSAB: Coarse-Grained Force Field for the Study of Protein–Protein Interactions and Conformational Sampling in Multiprotein Systems. *Journal of Chemical Theory and Computation*:acs.jctc.5b00660.

> P. Sfriso contributed to design research, analysed results and contributed to the writing of the paper.

Sfriso P, Emperador A, Gelpí J, Orozco M (2014) Discrete Molecular Dynamics: Foundations and Biomolecular Applications. in *Series in Computational Biophysics* (CRC Press), pp 339–362.

> P. Sfriso contributed with ideas and to the writing of the chapter.

Sfriso, P; Duran-Frigola, M; Mosca, R; Emperador, A; Aloy, P and Orozco, M. (2015) Residues coevolution guides the systematic identification of alternative functional conformations in proteins. *Structure*: *in Press.*

> P. Sfriso was the main developer of the project together with M Duran-Frigola and P. Sfriso contributed to the writing of the paper.

In preparation

Cunha, ES; Sfriso, P; Rojas, AL; Hospital, A; Orozco, M and Abrescia, NGA. The flexibility of the human cellular receptor CD81 large-extracellular-loop serves to enable Hepatitis C virus entry. *In Preparation.*

P.S ran the simulations, analysed the results and contribute to the writing of the paper.

Sfriso, P; Hospital, A; Buitrago, D; Mosca, R; Emperador, A; Gelpi, JL; Aloy, P and Orozco, M. TransAtlas: an integrative database of conformational transitions of proteins. *In Preparation.*

P. Sfriso is the main developer of the project together with A. Hospital and P. Sfriso contributed to the writing of the paper.

PhD Advisor

Prof. Dr. Modesto Orozco López

Chapter 1: Introduction

1.1 Proteins are Flexible Molecules

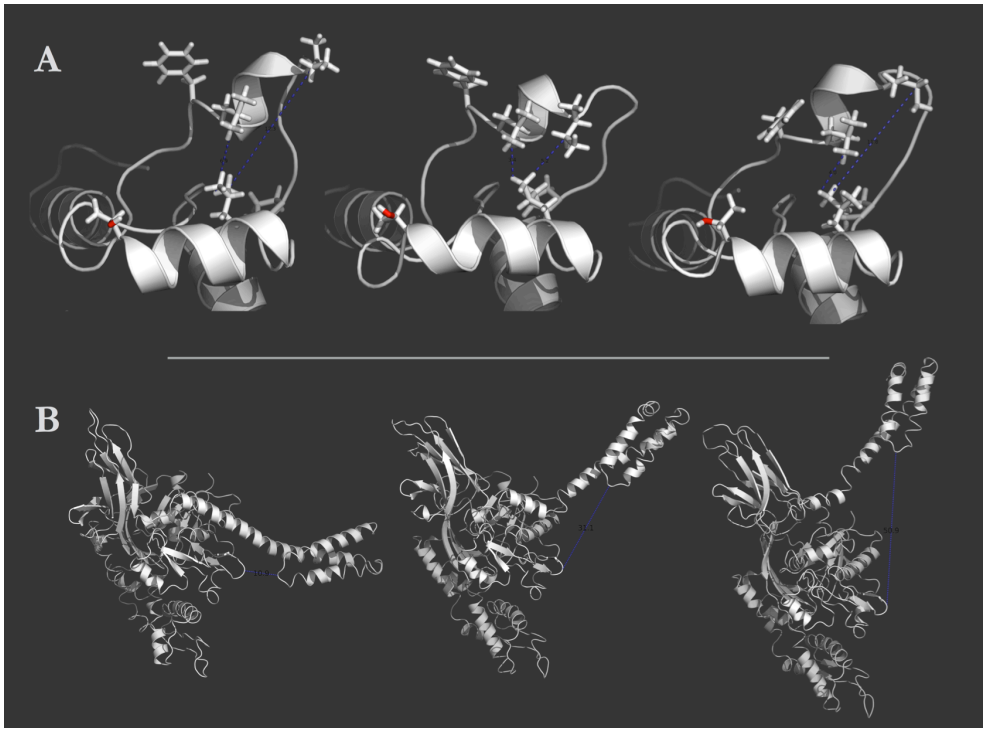
It is said that simpler is best. Science has been following this old, powerful, strategy for a long time now. However, in 1958 science received a nice surprise when Kendrew and Perutz solved the first structure of a protein –myoglobin. Why so complex? The myoglobin structure showed a fascinating spatial arrangement of atoms. A little baroque, yet precise, atomics interactions displayed an underlying rational that scientist rapidly struggled to understand. Why evolution shaped proteins in that manner? Proteins carry out a spectacular variety of functions that logically, demand equally diverse structural motives. But proteins are only made of roughly 20 amino acids, leaving no possibility other than complex structures. In one line, evolution chose to create complex structures with simple building blocks that facilitate the coding in the genetic alphabet.

Proteins form the scaffold supporting the cells, catalyse biochemical reactions, recognize ligands, sensor environmental changes, express genes, and lead the muscle contraction, just to name a few function examples. In other words, they are the machinery of the cell, responsible for living functions, and as any machine, they move to perform their task. Movement is crucial for protein functioning, and accordingly it has been refined during years of evolution, coded explicitly in the sequence (*1-4*). Protein motions can be subtle and concentrated in a small region of the protein (Figure 1A) but can also completely change the shape of the protein (*2*, *5*, *6*) (Figure 1B). In summary, we should scape from the classical view of proteins exhibiting well defined three dimensional structures and should move into the concept of conformational ensemble as the most realistic model to represent proteins (*2*, *7-9*).

The transformation between the most prevalent protein states in the conformational ensemble is known as conformational transition, and it is the main topic studied in this Thesis. Conformational transitions constantly occur in the cell and they are only possible by virtue of thermal motion. Thermal motion shakes structures providing the energy necessary to explore neighbour conformers, eventually hopping from one state to another. However, to ensure robustness, thermal motions are often coupled to external stimuli such as interactions with other molecules, chemical modifications, electrical impulses, mechanical stress, or any other kind of signal to trigger a conformational transition.

To understand protein flexibility and to decipher the complex mechanisms linking sequence to function has been, and it is, a central question in molecular biology that has attracted scientists from many fields: from pure theoretical to experimental biophysics. In this regard, there are several experimental techniques that contribute to understand protein motions. The first to mention is, X-ray crystallography, which provided with most of the structures deposited in the PDB database (*10*). In X-ray crystallography proteins are purified and crystallized, then, an X-ray beam diffracts on the crystal giving information of the 3D arrangement of atoms. It is clear from the procedure that conformational changes must be minimized to obtain comprehensive data, assuming that X-Ray structures are in many cases "static pictures". Nonetheless, insights of protein ensembles can be obtained from X-Ray in cases where different structures are obtained as consequence of varying crystallization conditions. The second most relevant high-resolution structural technique is Nuclear Magnetic Resonance (NMR)(*11*) that on contrary to X-ray crystallography, resolves structures in solution. Besides obtaining structures, NMR investigates protein plasticity, particularly dynamics time scales which is essential to set the bases for theoretical studies (*12-14*). Unfortunately, solution NMR is limited to medium to small protein size, which hinders the study of long conformational transitions in large systems. Other

Figure 1 Protein Motions



Proteins move in a wide range of motions. A) Local re-arrangements in the CD81-Large Extracellular Loop. B) Massive translocation of protein domains in the heat-shock protein Hsp70. While local motions (A) are within 2 Å RMSD large conformational transitions can lead to 40 Å RMSD displacements.

techniques showing protein dynamics are time-resolved crystallography (*15-17*), Förster resonance energy transfer (FRET) (*18*, *19*) and neutron scattering (*20*, *21*). Also, very recently Electron Microscopy (EM) has provided unprecedented detailed snapshots of large biomolecular machines (*22*), like ribosomes in action (*23*, *24*). However, despite the impressive advances in experimental methods, all these techniques are still unable to produce, in most cases, a sufficient high-resolution representation of protein movements. So, involving computational models seems inherent to the molecular flexibility issue. Molecular simulations mingle experimental observations with the underlying physic laws, and nowadays, they are the single technique able to dissect atomistic motions (either as stand-alone technique or in combinations with experiments).

1.2 Molecular Simulations: a computational microscope

Nearly 40 years ago computational biology was founded and started to operate as a *computational microscope* to observe macromolecules in motion (*25*, *26*). There are several simulations methods depending on the level of detail desired, although Molecular Dynamics (MD) is the reference one. MD appeared, in its discontinuous formulation, in the late 50s when theoretical physics ran studies on hard spheres interactions (*27*). From there, it took about 20 more years to extend MD to liquids and molecules (*28*), and finally in 1977 the first atomistic simulation of a small protein appeared (*29*). The "magic" of MD rests in its extraordinary intuitive algorithm that capitalizes how scientists think about molecular processes (for detailed description see (*30*, *31*)). It works as following:

i) Give particles an initial position $\overrightarrow{r_0}$ and initial velocities $\overrightarrow{v_0}$.

ii) Get forces and acceleration acting on the particles $\vec{F} = -\nabla V(\vec{r}(t))$ and $\vec{a} = \vec{F}/m$

iii) Integrate accelerations ($\vec{a}$) to obtain particle velocities ($\vec{v}(t)$).

iv) Move particles $\vec{r}(\mathrm{t} + \Delta\mathrm{t}) = \vec{r}(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{a}\Delta t^2$

v) Advance time $t = \mathrm{t} + \Delta\mathrm{t}$

vi) Repeat ii) to iv) as long as you need

Additional steps are required if simulation is expected to keep constant some macroscopic properties, such the pressure or the temperature. Note that the algorithm is simple thanks to energy function $V(\vec{r})$, which is indeed the key idea behind molecular simulation methods (*32-34*). $V(\vec{r})$ is referred as the force field: it dictates how the system moves, and defines the limits the conformational space that can be visited. The level of detail of the force field depends on the question that we want to address and it will be discussed in the Energy Description section. We cannot forget that the quality of any MD-derived result will be never superior to the quality of the force field used to derive the conformational ensemble.

The MD procedure is elegant and simple but suffers from severe intrinsic limitations. The most important one is that to follow biological motions (let say 1 second of a protein life), we typically need to repeat the algorithm 10^{15} times, since Δt is in the femtosecond time scale. Something far beyond of what can be simulated even in the largest super-computers. Numerical stability of computations demands integration steps smaller that fastest motion in the system. In biology, this points to chemical bond vibrations, happening in the femtosecond time scale. Unfortunately, it is not as easy as freezing vibrations from bods: in that case, time scale can be slightly larger, however limited to ~ 10 femtosecond regime from other movements.

Large scale motions, including the most dramatic conformational transition: folding a protein. Protein folding has been a permanent challenge for atomistic MD simulations. In principle, for an infinite computing capability and for a highly accurate force field a random conformation of the polypeptide chain, placed in a physiological-like environment should spontaneously fold into the native conformation. However, it took to the scientific community more than 30 years to systematically fold based on physical principles a selected set of small proteins (*35*), showing very fast folding kinetics. This milestone in scientific computation required the design of a special purpose computer, besides from many refinements of force field parameters and algorithms heuristics over three decades. But still, even the fastest computers and the most accurate force fields are still unable to fold complex proteins. Thus, despite its undeniable success, atomistic simulations of large conformational movements are still very inefficient, due mostly to the problems for sampling accurately the protein conformational space.

1.3 The Sampling Problem

In this section I will describe the interest behind sampling the whole conformational space for a protein, and the problems associated to this idea.

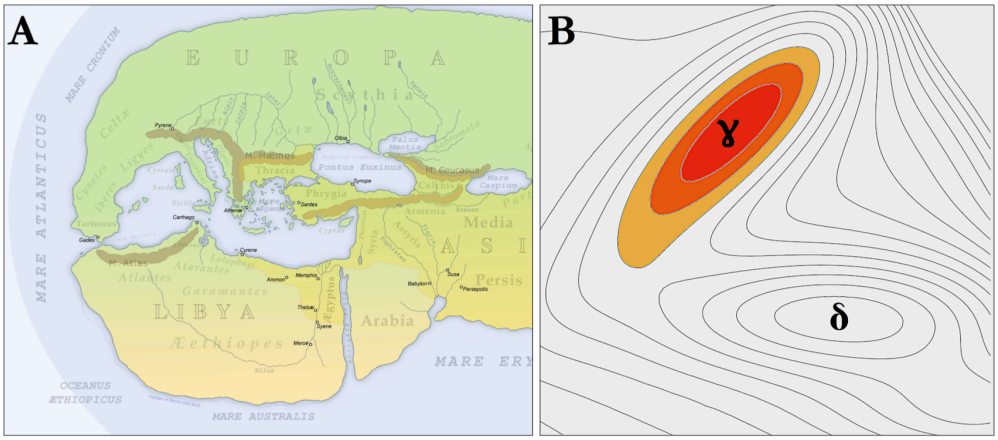
In statistical mechanics, the partition function (Z) of any system (a protein and its environment in this case) is defined, in the discrete formulation, at constant temperature and fixed volume, as:

$$\text{(Eq. 1)}\quad Z = \sum_j^N e^{-\frac{E_j}{k_b T}}$$

where E_j is the energy for j microstate, k_b is Boltzmann's constant, and T is the absolute temperature. E_j depends on the coordinates and momentum of every atom of the system, including the solvent. N is the total number of microstates of the system. At thermodynamic equilibrium, the partition function fully captures the properties of a system. Hence, most thermodynamic variables, such as free energy or entropy can be derived from the partition function or its derivatives. Accordingly, after all possible N-microstates are visited in a simulation, the free energy of any conformational transition can be exactly computed. But, for that purpose, equilibrium sampling requires access to all regions of conformational space or at

least, those regions with significant population (*36*). The technical impossibility of accessing all accessible microstates is known as the sampling problem.

Figure 2: Sampling problems over history



(A) Herodotus map of the world in fifth century BC. Herodotus map was completely biased towards the Mediterranean Sea, where he lived. The lack of fast boats made impossible a good representation of the world. (B) Poor sampling situation in a protein conformational landscape. Sampled structures (colored) are close to the initial one (γ) and other relevant structures are never visited (δ). In both cases, poor sampling impedes reasonable predictions away from the initial point.

Going back to molecular simulations, there are two main effects that preclude the exhaustive sampling of the conformational space (given its high dimensionality). The first one, as mentioned, is the very short integration step, and the second one is the choice of the initial coordinates (initial structure) to start a trajectory. Initial structures are obtained from experiments and consequently there are stable conformers, constituting kinetic traps. Let me use a sailing analogy to explain the consequences of biased sampling. In ancient Greek times, Herodotus recollected the available information and elaborated a map of the world (Figure 2A). It easily seen that Herodotus lived in what is nowadays the Turkey area because in those times shore geology was explored –sampled– by sailing. Their sailing capabilities where rather limited: Herodotus world map is only acceptable in the Mediterranean Sea, despite of centuries of sailing adventures. Similar situations are faced in molecular simulations daily (Figure 2B), where the starting (γ) conformation usually dictates the accessible conformational space (in a finite computation time). Spontaneous large conformational explorations (i.e. American continent, or δ state in Figure 2) are, therefore, unlikely. The implications of protein-sampling problem worsen considering that there are not *satellite maps for proteins*, so there is no reliable way to detect missing information.

1.4 Optimizing Molecular Simulations

There is no discussion that after more than 30 years of atomistic simulations the equilibrium sampling is still an unsolved issue (*36*). Waiting for better software and hardware will not solve the problem, and strategies to improve predictive power of simulations are required. This section concentrates the efforts to optimize sampling capabilities, beyond waiting for longer trajectories. The elements discussed here range from developing alternative 1) sampling methods, adjust the 2) energy description and reduce the degrees of freedom of the 3) protein and 4) solvent particles.

1.4.1 Sampling Methods

Undoubtedly, MD is the reference sampling method in computational biology. The MD algorithm was described previously as part of a more comprehensive contextualization. It is MD limitations in accuracy and efficiency that propelled the search of alternative methods. In terms of its accuracy, it can be outperformed by Car-Parrinelo MD (*37*) and Born-Oppenheimer MD, that are reviewed in references (*38*, *39*), while here I will centre the discussion on methods trying to overcome MD sampling (efficiency) limitations.

The first rational explored is to disregard the temporal connection between states, which most of the times is not essential. This is the case for Monte Carlo simulations.

Monte Carlo

Scientists working on the atomic bomb developed the Monte Carlo (MC) sampling method in the 1940s. Instead of using the potential energy of the system to propose new conformations, MC tries new conformations randomly and evaluates the energy afterwards. In biomolecular simulations, conformations are obtained by trial changes in dihedral angles (or any internal coordinate). This allows MC to propose, in a natural manner, reasonable conformations waiving the complexity of moving in Cartesian space. Note that MC can handle constraints such as fixed bond lengths by restricting the sampling over that coordinate. The choice to use MC is usually for convenience: it enables the use of any energy interaction either complex or very simple like step functions. The main disadvantage is, as mentioned, the lost of the temporal connection between states and the intrinsic problems related to the reduction of the degrees of freedom. MC it is used in non-statistical approaches such as protein-protein and protein-ligand docking and it also powering folding software as ROSETTA (*40*). A nice example of MC was presented by Ding et al for peptides (*41*), that I will use to illustrate its sampling power. Ding et al showed that when libraries of amino acid configurations were computed in advance, swapping residues conformers as a MC try caused fast motions in the conformational space. In light of this, MC is a powerful sampling engine when some intuition can be used to select the internal coordinates for

example in the study of conformational transitions (*42-46*).

Normal Mode Analysis

Analysing directly the energy surface can reveal equilibrium fluctuations encoded in the structure. Normal Mode Analysis (NMA) extracts that information directly from the force field, without running trajectories (*47-51*). NMA assumes that the initial structure is in a deep energy minimum and its conformational energy can be approximated as a multidimensional parabola (*49-51*). Consequently, at the energy minimum, the potential energy (V) can be represented as a harmonic-truncated Taylor series, characterizing the energy function only by its second derivative (since the first term vanishes). Given a structure with N particles (without solvent):

$$\text{(Eq. 2)}\quad V(r) = \frac{1}{2}\sum_{i,j}^{3N} \frac{\partial^2 V}{\partial r_i\, \partial r_j}\, \Delta r_i\, \Delta r_j$$

where r_i are the coordinates for the i[th] atom and Δr_i its relative displacement (identical for j index). The second derivative term in Eq. 2 is the element i-j of the *Hessian* matrix of the system. Eigenvectors and eigenvalues of the Hessian, weighted by particles mass, lead to 3N-6 concerted displacements referred as normal modes. In NMA, it is assumed that normal modes with the largest fluctuation are the functionally relevant ones: like function, they exist by evolutionary design rather than by chance (*4*). Although there is evidence of the breakdown of the harmonic hypothesis (*52*), NMA has become very popular in the field of protein dynamics. It can be applied to all resolutions including the atomic one (giving that the second derivative of the force field is well-defined), but is usually coupled to coarse-grained and elastic network models (see Energy Descriptions) (*48*, *53*, *54*). An original extension of NMA was presented by Bathe, where normal modes are computed using protein shapes rather than atomic coordinates (*55*). Bathe obtained intrinsic motions of two model proteins based only on the solved-excluded volume, being a nice complement to Electron Microscopy technique. To sum up, traditional NMA consists in three stages. 1) Minimize the initial structure in the selected force field, 2) compute the Hessian matrix and 3) obtain the normal modes by diagonalization of the Hessian matrix. From there, equilibrium fluctuations are disclosed in good agreement with more advanced simulation techniques (*48*).

Discrete Molecular Dynamics

Discrete Molecular Dynamics (dMD) is an inexpensive sampling engine alternative to MD. It works with any possible protein representation, but for its capabilities, dMD seems tailored to work in conjunction with coarse-grain representations of the proteins. dMD is the main sampling method used in this Thesis, so, I will stress the algorithm particularities and applications. However, I will not detail the algorithm formalism in this introduction because it can be found

either in Chapter 3 of this Thesis or in methodological reviews (*31*, *56*, *57*).

Discrete Molecular Dynamics was indeed the first formulation of MD (*27*), designed as a proof of concept method for hard-spheres simulations. Karplus and co-workers restored it back for biology in late 90s, again, needing a fast proof of concept tool to investigate folding thermodynamics (*58*, *59*). Since then, dMD has been constantly expanding it applications being outstanding in protein and RNA folding (*60-64*). dMD also excelled in reproducing protein flexibility (*48*, *65*), specially in protein-protein docking poses (*66*). It was also applied in the relaxation of homology models (*67*), small ligand screening (*68*) or protein aggregation (*69-72*).

The dMD algorithm avoids computing forces by assuming that particles move in the ballistic regime. In other words, particles move with constant velocity until an event (collision) occurs, making the trajectory advance event-wise instead of fixed step-wise (*64*, *65*). Imposing the energy (Eq. 3) and momentum (Eq. 4) conservation rules, the temporal evolution of the system is followed. This is the case for a collision between particles i-j:

$$\text{(Eq. 3)}\quad \frac{1}{2}\, m_j\, {v_j}^2 + \frac{1}{2}\, m_i {v_i}^2 = \frac{1}{2}\, m_j\, {v_j}'^2 + \frac{1}{2}\, m_i\, {v_i}'^2 + \Delta V$$

$$\text{(Eq. 4)}\quad m_j\, v_j + m_i v_i = m_j\, v_j' + m_i\, v_i'$$

where m_j is the mass of particle j, v_j and v_j' are the velocities of particle j before and after colliding. ΔV is a potential energy term to allow inelastic collisions. From Eq. 3 and Eq. 4 the velocities after the collision are obtained. Finding collision times requires most of the computational time, since other steps are elementary. A schematic dMD algorithm is presented below:

i) Give particles an initial position $\overrightarrow{r_0}$ and initial velocities $\overrightarrow{v_0}$.

ii) Compute collision times (τ_{ij}) for all i-j interacting particles

iii) Find the minimum collision time $\Delta t = \min\{\tau_{ij}\ \forall\ i,j\}$

iv) Move particles $\vec{r}(\mathrm{t} + \Delta\mathrm{t}) = \vec{r}(t) + \vec{v}\Delta t$

v) Advance time $t = \mathrm{t} + \Delta\mathrm{t}$

vi) Update i-j velocities according to ballistic equations of motion (Eq 3 and 4. Note that this step is only necessary for colliding particles).

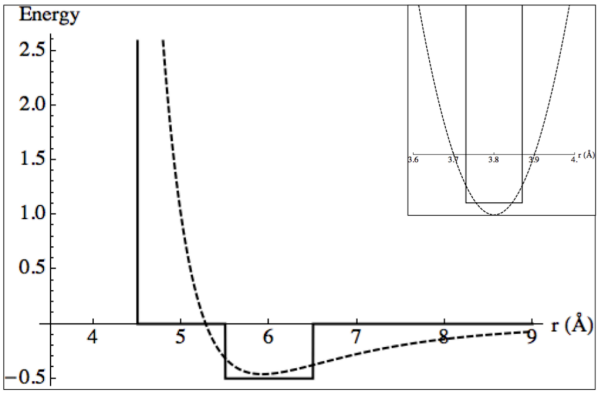
vii) Update collision times τ_{ij}

viii) Repeat ii) to vii) as long as you need

The dMD framework requires the use discontinue potentials (Figure 3), with flat energy minima and energy discontinuities at interaction distances. The number of discontinuities in the interaction potential depends on the resolution desired, but usually 2-3 steps are used since the computational cost increases linearly with the number of them (*73*). In the limit of infinite

discontinuities, dMD trajectories are identical to MD ones, but extremely inefficient. Although convenient, the use of step potentials introduces its own problems. First, the algorithm is more complex, since dealing with large numbers of collision events in an efficient manner requires careful attention to data organization. The second problem is memory handling: storing the information describing events it is far from trivial (*31*, *73*). The third problem is that event executions can cause next events to be invalidated and inserted anywhere in the queues of collisions, making parallelization of dMD codes a rather challenging task (*74*).

Figure 3: dMD square potentials



In discrete Molecular Dynamics continuous energy functions (dashed) are replaced with step potentials with flat energy minima. A typical non-bonded vdW interaction and a bonded interaction (inset) are represented. At infinite discontinuities both energy potentials become equal. Energy is in arbitrary units.

dMD is an appealing sampling engine for many reasons besides its efficiency. For instance, it can accommodate easily any modification of the energy landscape, like spatial restriction or modifications in the energy interaction function, convenient to define multiple minima profiles. Finally, dMD can handle naturally uncertain interactions, and is very robust to structural errors (*67*), both properties that were crucial for the objectives of this Thesis.

1.4.2 Energy Description

The sampling engines so far described survey the energy landscape to find the accessible regions of the system and, ideally, their free energy. However, the energy landscape definition varies with the level of resolution, which is in turn dictated by the biological question. The different energy descriptions commonly used for proteins are presented here under four main categories: quantum, classical, coarse-grained and elastic network models.

Quantum

Theoretically, dynamics of any molecule (nuclei and electrons) can be studied by solving Schrödinger time dependent equation, although there is no known exact solution for any molecule. From there on, approximations are needed to model molecular motions. The first one is to use the solutions of the time-independent Schrödinger equation (Eq. 5) as a base for

molecular wave functions:

$$\text{(Eq. 5)}\quad H\,\psi(r) = E\,\psi(r)$$

where E is the energy and $\psi(r)$ is the molecular wave function. H is the Hamiltonian operator. An example of Hamiltonian defined for a single non-relativistic particle, moving in an electric field, is:

$$\text{(Eq. 6)}\quad H = \frac{-\hbar^2}{2\mu}\nabla^2 + U$$

where $\hbar$ is the Planck's constant (divided by 2π), μ is the reduced mass of the particle, U its potential energy and ∇^2 is the Laplacian operator. Hamiltonian operator gives the total energy of the system (the wave function), in this case as the sum of the kinetic and potential energy. One of the main advantages of QM approach is that "in principle" energies can be computed as accurately as desired. Expectedly, the problem is again its computational cost, limiting QM methods to small isolated molecules. Regarding protein dynamics, the quantum approach implies the use of very simple QM descriptions (semi-empirical Hamiltonians or low level DFT calculations) or hybrid approaches whether electronic effects are considered only in specific areas such as the catalytic site, modelling the rest of the system in a classical fashion (QM/MM methods).

Classical

The classical approach implies ignoring electrons degrees of freedom; instead their effect is captured in a set of parameters derived in a semi-empirical way. This approach greatly simplifies equations used to represent the molecular Hamiltonian, namely the force field. The force field is parameterized to reproduce experimental observables, or high-level reference quantum-mechanics calculations. Force fields underwent into extensive refinement in their parameters, but the basic formalism has barely changed from its origins (*32*, *34*, *75*) and is common to all variants (Eq. 7). It can be schematically represented as:

$$\text{(Eq. 7)}\quad H = H_{bonds} + H_{angles} + H_{torsions} + H_{electrostatic} + H_{vdW}$$

In this formalism, atoms are treated as spheres, while molecular stereochemistry is ensured connecting neighbouring atoms through springs (Eq. 7a and 7b). Dihedral angles are dictated by periodic functions (Eq. 7c) that due to their critical importance on the simulation outcome have been finely characterized. In the same line, long-range interactions arising from atomic charges or from van der Waals interactions are modelled with simple classical potentials (Eq. 7d and 7e, respectively). Typical classical-atomistic force field terms are herein presented (note that all

energetic contributions solely depend on the atoms position):

(Eq. 7a) $H_{bonds} = \sum_{bonds} K_s\ (l - l_0)^2$

(Eq. 7b) $H_{angles} = \sum_{angles} K_b\ (\alpha - \alpha_0)^2$

Where K stands for the stiffness of the bond stretching (K_S) or bond bending (K_b). Bonds (l) and angles (α) equilibrium positions are defined used a reference value, indicated with 0 sub-index.

(Eq. 7c) $H_{torsions} = \sum_{torsions} \frac{1}{2} V_t\ [1 - \cos(n\varphi - \delta)]$

V_t is the torsional barrier, φ the torsional angle while n and δ are accounted for periodicity and phase of the energy function.

(Eq. 7d) $H_{electrostatic} = \sum_{charges} \frac{Q_i Q_j}{R_{ij}}$

(Eq. 7e) $H_{vdW} = \sum_{nonbonded} \epsilon_{ij} [\left(\frac{A_{ij}}{R_{ij}}\right)^{12} - 2\left(\frac{C_{ij}}{R_{ij}}\right)^{6}]$

Q represents atomic charges, R distances and A, C and ϵ correspond to parameters from Lennard-Jones potentials. Finally, it has to be said, that despite the simplistic nature of the force field, their accuracy have proven striking in many cases, being very reliable for standard calculations (*76*). Further interested readers are referred to reference (*38*) for an extended review.

Classical force fields meet their limit when simulating chemical reactivity. In those cases combined quantum-mechanics/molecular-mechanics (QM/MM) methods are the most promising choice (*39*, *77*, *78*). QM/MM uses CPU-demanding quantum formalism only in critical points of the protein, such as the catalytic site, and describes the flexibility of the rest of the protein classically. The boundary area must be modelled sensibly, being dummy atoms strategies or localized orbitals are the most popular ones. Eq. 8 describes the total energy contribution in the QM/MM approach:

(Eq. 8) $H_{total} = H_{QM} + H_{MM} + H_{QM/MM}$

In QM/MM formalism, wave functions of the QM region feel the rest of the system through the QM/MM region, since the solution of both of them are coupled in variational methods.

Other approach is Arieh Warshel's Empirical Valence Bond theory that describes enzymatic reactivity without explicit consideration of the electronic degrees of freedom (*79*). Instead, a library of chemical structures is computed at the QM level, used to represent the reactant entities. Resonance states of those structures are later incorporated in simple empirical potentials in

molecular mechanics calculations (*80*). This elegant approach has been largely successful, and gave many insights of protein reactivity. Finally, solely mention that parameterization of new reactive structures requires careful attention.

Coarse-Grained

Replacing groups of individual atoms with one particle in a lower resolution, coarse-grained (CG), representation enables the simulation of large-scale biomolecular processes. Five reasons make emerge CG modelling (*81*):

- To simulate huge systems containing millions of atoms
- To increase by 1000-fold the accessible time-scale, allowing the simulation of slow processes
- To facilitate high-throughput studies
- To smooth landscapes showing where details matter
- To provide with a computational inexpensive tool for test purposes

CG models can be derived in two ways. The first one needs a reference structure to build the force field, called "bottom-up" approach. Parameterization of bottom-up approaches uses reference atomistic simulations, which yielded so far the best CG accuracy. On the opposite side, "top-down" approaches are only built from the physical properties of particles. Their calibration uses thermodynamic data rather than reference simulations, therefore being transferable to other systems. The balance is delicate: transferable models are needed to simulate extremely flexible molecules but their accuracy is not ideal. Although this classification is useful to discern applicability, it is becoming less clear as mixed force fields are appearing.

A second common classification among CG models, related to the previous one, distinguishes between physics based and knowledge based approaches. Physics based approaches infer, using physical theories, the interactions in CG models, either via bottom-up strategies or via top-down strategies. In contrast, knowledge based models are constructed on the basis of information extracted from either one reference structure (structure based) or from a collection of experimentally determined structures (knowledge based).

Physics Based Coarse Grained Potentials

Derived from the underlying chemistry, either formally or by intuition, physics-based models retain chemical specificity of particles. The level of resolution depends on the biological problem: detailed backbone representations are needed to model secondary structure changes while side chain detail is critical to protein interactions. There are several successful, well-established, approaches among them MARTINI and UNRES force fields and two very promising emerging ones: PaLaCe and PRIMO.

- MARTINI force field was extended for proteins after the success of the version for lipids, from which inherited the 4 to 1 mapping of atoms to CG sites (*82*). Amino acids are represented with one bead for the backbone and none to 4 (Trp) beads per side chain. An elastic network model (see below) is used to preserve the secondary structure of the reference structure, therefore being structure dependent (*83*). In MARTINI force field particles are classified in four categories according to their chemical nature in: charged, polar, non-polar and apolar. Each category is then split in sub-types giving a total of 20 bead types. MARTINI became one of the top used CG force field being implemented in all major molecular dynamics simulation packages. It is widely used to simulate protein-membrane interactions and protein supramolecular arrangements. However, due to the secondary structure constrains it can not be used to simulate folding nor aggregation processes

- UNRES model describes the backbone with two CG sites and side chain with a single ellipsoidal site (*84*). UNRES has been constantly improving to become one of the most transferable CG force fields (*85*). Bonded terms (bonds, angles and dihedrals) are defined for backbone particles and a rotational potential is used to represent side chains preferred rotamers. In the latest version, all non-bonded terms are obtained from *ab initio* calculations of small systems and from Potential Mean Forces extracted from atomistic MD simulations. UNRES has been largely used to study protein folding or other structure prediction problems being able to reasonably predict (RMSD 3.5 and 5.5 Å) two native structures in a CASP exercise (*86*, *87*). One disadvantage of UNRES model is the difficulty to put back all the atomic detail for multiscale modelling.

- PaLaCe force field was developed to investigate the mechanical properties of proteins (*88*). It represents amino acids with two types of beads: the first class guarding for an accurate dihedral description and correct hydrogen bond geometries while the second class deals with non-bonded interactions. Combination of both of

them leads to an atomistic representation of the backbone and CG side chains resolution, up to three beads. Despite the high number of particles, PaLaCe achieves a 1000 fold increase in efficiency compared to atomistic simulations, since only a subset degrees of freedom are followed every step. PaLaCe was parameterized in a bottom up approach with the Iterative Boltzmann Inversion method (*89*), applying restrains for secondary structure, which limits its current applicability to reproduce large-scale dynamics. The mechanical properties of the protein are accurately captured in pulling experiment coinciding with Atomic Force Microscopy data. All together, this is a very promising approach to simulate processes where folding/unfolding does not occur.

- PRIMO is a physics-based force field tuned for optimal description of backbone and side chain dihedrals (*90*). PRIMO energy functions are based on the CHARMM force field, ensuring transferability. Amino acids are modelled with three beads for the backbone and one to five beads for side chains. The main advantage of using PRIMO force field is that its CG sites are carefully placed to allow an analytical reconstruction to atomistic representation. This scenario is ideal for innovative multiscale approaches that will be protagonist of next decade simulations methods.

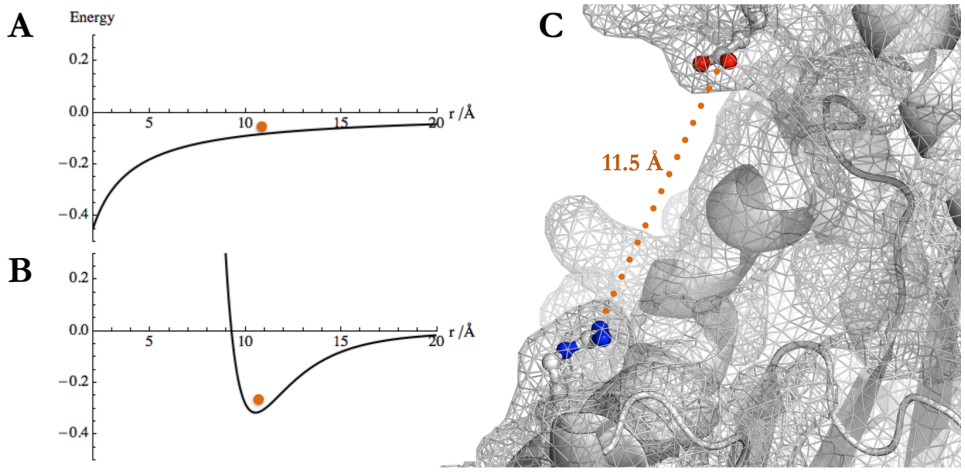
In a more detailed representation, Medusa force field (*91*) combines an implicit solvent model (*92*) with a physical approach, specially suitable to treat ligand-protein interactions (*93*). Other important CG contributions are SIRAH (*94*), OPEP (*95*, *96*), SCORPION (*97*), the Bereau and Desderno Model (*98*), MS-CG by the Voth Lab (*99-101*) are not detailed here for space limitation, despite their relevance and popularity.

Structure Based Potentials

Provided that high-resolution structure exists, protein dynamics can be rigorously modelled solely based on the 3D position of its atoms. This approach is referred as native-centric, Go-like or structure based models (SBM). Pioneering the field, Go and co-workers observed that protein motions are consistent with the folded structure of native state (*50*, *102*, *103*), considering that the strength of non-bonded interactions is determined by the folded structure, rather than by the physicochemical identity of the residues. The approach is supported from the *minimal frustration* principle (*104*, *105*) in the sense that the native state not only optimizes the stability entire protein, but also optimizes all individual interactions (see below).

In practical terms, SBM models are expressed in in terms of energy potentials that place the energy minimum at the native structure. In the simplest SBM, each residue is modelled with a single bead centred in Cα position and attractive Lennard-Jones potentials are used for native contacts. Non-native interactions are modelled with repulsive potentials and bonded interactions are treated like springs, again centred at reference distance. In spite of their extreme simplicity, SBM consider excluded volume, describing accurately the entropy configurational loss as the protein folds. Moreover, by eliminating the frustration of non-native interactions the energy landscape becomes ideally funnelled predicting folding pathways (*58*, *106*) or folding kinetic rates (*107-109*). Also SBM have been largely used to model conformational transitions in protein with great success (*110-112*). In conclusion, these intriguing results suggest that topological native contacts, regardless their nature, can explain a large fraction of protein dynamics (*113*, *114*).

Figure 4: Approaches to model residues interactions



In this example, the two main ways of modeling residue interactions are presented. In Adenylate Kinase (PDB 4AKE), an aspartic acid (negatively charged) 11.5 Å away interacts with an arginine (positively charged). In physical modeling (A) the particles interact through Coulomb law therefore they will attract each other until are at collision distance. Instead, in structure-based model (B) they attract each other to reference distance in (C) (11.5 Å, orange circle) and not closer than that. Note that the minimum position in the potential energy interactions for each case is different. Energy has arbitrary units.

In summary, SBM do not consider the physical properties of the residues just their position in the 3D structure of the structure, being not transferable to other macromolecule. However, SBM are the most successful models to describe protein fluctuations and folding pathways and it is expected that as the number of high-resolution structures grows SBM will be increasingly used.

Knowledge based Potentials

Contrary to SBM where generally one structure is used to build the energy potential, knowledge based potentials collect structural information from an ensemble of high-resolution proteins. Averaging over many structures yields more transferable potentials, especially for globins, similarly to SBM, aiming to place the energetic minimum at the folded structure (*115*). Knowledge based potentials are also known as statistical potentials because interaction functions are based upon the frequency of distances in the reference ensemble of structures. Furthermore, the distribution of residues distances reflects also the interaction with the environment, solvent included. In this regard, Miyazawa and Jernigan showed in a seminal work that solvent effect could be adopted by means of statistical potentials (*116*, *117*). They also showed that native pair potentials are minimized for folded structures, keeping non-native contacts less favourable. Knowledge based potentials found major applications in structural prediction by Homology Modelling and in the validation of protein models in software packages like PROSA (*118*). Also, knowledge-based backbone dihedrals are coupled to physical potentials to better describe the secondary structure propensity of peptides chains (*64*). David Baker's ROSETTA software uses library of structural fragments based on statistical potentials being the leading approach in *ab initio* protein structural prediction (*119-122*). One obvious drawback of knowledge based approaches is their incapacity to make predictions where there is not much known, like intrinsically disordered proteins dynamics.

Elastic Network Potentials

Elastic Network Models (ENM) are conceptually linked to the Go-like models and implicitly assume the minimum frustration paradigm. ENMs were designed to describe equilibrium oscillations, representing a clever shortcut to sample equilibrium conformations (*123*). Similar to SBM, ENM are parameterized used a reference structure, whose deformation energy follows a harmonic functional, in Cartesian space:

(Eq. 9a) $$H = \sum_{i,j} \delta_{ij} K_{ij} (r_{ij} - r_{ij}^0)^2$$

(Eq. 9b) $$\delta_{ij} = \begin{cases} 1 & if\ r_{ij}^0 \leq Rc \\ 0 & if\ r_{ij}^0 > Rc \end{cases}$$

where i and j are particles (atoms or CG sites), K_{ij} is a spring constant or a simple function of internal distance (*53*, *124*). r_{ij} stands for inter-residue distance and the superscript 0 refers to the value of r_{ij}^0 in the reference structure. Rc is the maximum distance for a contact to be considered native. Despite employing only two parameters (constant K_{ij} model) ENM can accurately reproduce protein equilibrium fluctuations. When low-resolution (1 bead) ENM models are coupled to NMA, the computational cost is minimal but the accuracy is maintained. Somehow, this shows that the harmonic approximation of NMA is coarser than the ENM protein description. ENM-NMA showed valuable estimations in biological processes including conformational transitions, protein-protein docking (*125-127*), protein-ligand docking (*128*) or even to relate disease mutations with dynamics (*129*). It is particularly remarkable, the success of ENM-NMA in finding transitions paths, showing that extrapolation of equilibrium motions of the initial structure unveils accurate path predictions (*130-137*). One drawback of ENM is that is performed in Cartesian space is that bond distances or angles are distorted at a given amplitude of motion, but it can be avoided switching to internal coordinates (dihedral angles), automatically fixing covalent geometry (*138*).

1.4.3 Protein Representation

When constructing a particle-based model for a protein, the first step is to define the particles used to represent the system (atoms in case of atomistic simulations). Do we need to consider electrons explicitly? Can we use a low-resolution efficient model? The universal protein model does not exist. In other words, protein resolution must be rationally adjusted to the biological question to maximize sampling quality. Although this is a trivial statement, it is often ignored while using standard or automated protocols. Optimized protein model must:

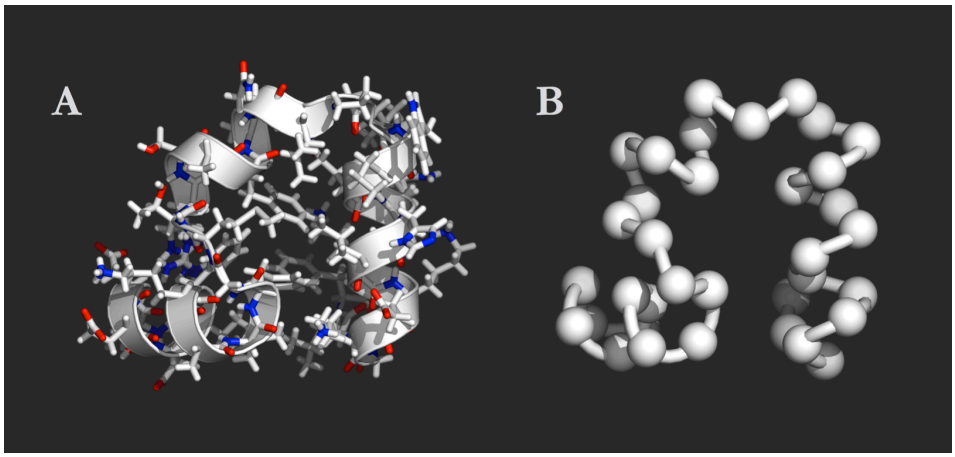
- Preserve the features needed to describe the phenomena of interest, for example, electronic details for enzymatic reactions or hydrogen atoms to accurate hydrogen bonds reproduction.
- Eliminate or average out sufficient detail to ensure feasibility
- Provide with an understandable description of governing physical forces
- Allow for a smooth change of resolution if needed.

There is a great range of choice for protein resolution: from electronic description to proteins modelled as rigid entity. For the purpose of this Thesis four resolution levels are relevant:

- **Electronic.** In practice electrons degrees of freedom are in proteins are explicitly consider only at critical points, particularly the active site.
- **Atomistic.** In molecular simulations it is prevalent to use the atomistic representation where individual atoms are explicitly modelled without considering electrons distributions' degree of freedom. When affordable these models yield the best results thanks to the highly optimized force fields.
- **Pseudo Atomistic.** (also united-atom) This representation includes all atoms explicitly but hydrogen that is fused to its bonding heavy atom. This strategy already represents a significant speed-up for integration-based methods, where integration steps can be larger after fastest bond oscillations (involving hydrogen) are removed.
- **Coarse-Grained.** Grouping atoms together leads to CG representations (Figure 5). There are several CG strategies but almost all of them include a particle cantered at the Cα position that permits better backbone reconstruction. One bead per residue model is widely used in combination with SBM and ENM sampling approaches, providing results of surprising quality.

Although independent, protein resolution and energy description are often leveraged with few exceptions (*99*, *139*, *140*). This fact is summarized in Table 1 together the broad range of protein resolutions, and a collection of examples of applications for each model. For extended review of CG applications see references (*141-143*).

Figure 5: Protein Resolution



(A) Atomistic representation of the Villin head domain (PDB 1WY3), hydrogen atoms are excluded for simplicity. (B) Coarse-grained representation of the same protein. Residues are modeled with a single bead centered in the Ca position. Although residue interactions become solely topological, protein main features are easily recognized.

Table 1: Adjusting protein resolution to the biological process

Protein Resolution	Application example	Sampling method	Transfe-rable	Energy description	References
Atomistic	Binding process	MD	Y	Physics-based	(*144*, *145*)
	Enzymatic Reactivity	QM/MM	Y	Physics-based	(*146-149*)
	Folding	MD	Y	Physics-based	(*150-152*)
	Conformational transitions	MD	Y	Physics-based	(*45*, *153-155*)
	Protein Engineering	MC	Y	Physics-based	(*156-158*)
United-atom	Membrane interactions	MD	Y	Physics-based	(*159*)
	Protein Flexibility	dMD	Y	Physics-based	(*66*)
	Conformational Transitions	dMD	Y	Physics-based	(*160*)
	Folding	dMD	Y	Physics-based	(*62*, *63*, *87*)
6-4 beds	Aggregation	dMD	N	Ad-hoc	(*69*, *71*, *72*)
	Folding	MC/MD	Y	Knowledge/ Physics-based	(*161-163*)
	Membrane interactions	MD	Y	Physics-based	(*82*, *164*)
	Conformational transitions	MD	Y	Physics-based	(*165*)
2 beads per residue	Folding	dMD	N	SBM	(*58*, *59*)
1bead per residue	Protein Dynamics	NMA	N	ENM	(*53*, *133*)
	Protein Dimerization	MD	N	SBM	(*166*, *167*)
	Folding	MD	N	SBM	(*168*)
	Conformational Transitions	MD dMD NMA	Y	SBM/ENM	(*112*, *129*)
	Protein Design	MD	Y	Knowledge based	(*169*)
Domains as rigid parts	Protein Binding	MC	N	Physics-based	(*170*)
	Supra-molecular assemblies	BD	N	SBM	(*143*, *171*)
1 bead per protein	Protein-Membrane interactions	BD	Y	Physics-based	(*172*)
	Protein Diffusion	BD	Y	Physics-based	(*173-175*)

Table 1 shows the coupling between resolution, sampling method and energy description. BD: Brownian Dynamics. dMD: discrete Molecular Dynamics. MC: Monte Carlo. MD: Molecular Dynamics. NMA: Normal Mode Analysis. QM/MM: hybrid Quantum Mechanics-Molecular Mechanics. Physics based potentials refer to energy functions based on physico-chemical properties of particles. SBM: structure based models. Knowledge Based refer to energy interactions derived from the statistical analysis of known proteins. ENM: Elastic Network Models. Ad-hoc refers to energy interactions adjusted for a particular system and simulation. An energy description is transferable when parameters only depend on the intrinsic nature of the particle, therefore are valid for any system.

1.4.4 Solvent Representation

Water shows, despite its simple structure, complex collective behaviour, which is key to understand biomolecules. Water interacts with proteins in several ways: it screens electrostatics, reinforces hydrophobic interactions and forms hydrogen bonds in the surface. To capture water peculiarities in a force field is a challenging task reflected in the number of quite different strategies. The number of particles used to represent water molecule is a matter of debate. Highest resolution models use 5 sites to describe water molecules (one per each atom and 2 extra for lonely electrons pairs) being TIP5P by Jorgensen group is the most used one of this type (*176*, *177*). Using a dummy particle, 4 sites water molecules exist (*178-180*), giving better electrostatics distribution. However, the 3 particles water models, in particular TIP3P (*178*, *181*) and SPC/E (*181*), are by far the most popular ones due to their compromise between accuracy and efficiency.

At lower resolution, CG water models can be implicit or explicit. Implicit models are a very effective way to reduce the degree of freedom by simulating only the protein structure. In this regard, water effects (hydrophobic attraction and charge screening) are mimicked using two principal strategies: altering non-bonded potentials or adding an extra term in the force field (*182-184*). The idea is to represent the solvent as a continuous medium seems a natural way of improving the computational speed, although there are several open issues, for instance computing solvent entropy, See references (*81*, *183*, *185*, *186*) for details.

Explicit CG models of water can be obtained by fitting atomistic simulations or experimental water properties (*187*). In the first case, 1 particle per water molecule is the most popular representation when techniques such as Iterative Boltzmann Inversion (*89*) or Force Matching (*99*, *100*) are used for parameterization purposes. In the second case, the typical mapping is 1 particle per 4 water molecules, like in MARTINI force field (*188*, *189*). Experimental data of water density, diffusion rates, solvation free energies and water-air surface tension are used to shape water interactions into Lennard-Jones or Morse potentials (*190-192*). The lack of charges in mentioned CG water models impedes electrostatic screening, something that can be alleviated with implicit charge screening or polarizable water models. Instead of one big particle groping water molecules polarizable CG water molecules use 2-3 particles with explicit charge separation (*188*, *193-195*). In an effort towards multiscale simulations, the Pantano group proposed the WT4 CG water with 4 particles in a tetrahedral organization mapped to 11 regular waters (*196*). All four particles carry explicit charge and are connected between them with harmonic springs.

1.5 Algorithms to enhance sampling

Simulating biological processes requires a delicate trade-off between proper detail and affordable efficiency but even when balanced it can still be prohibitively slow. Extensive sampling is basic to study phenomena like activated processes. When the initial conformation constitutes a kinetic

trap, or large diffusive processes spanning very distinct protein conformers occur, then exploring the conformational space becomes very inefficient. In such situations elegant computational heuristics are applied to overcome sampling limitations (*197*, *198*). The main idea behind these techniques, generally referred as Enhanced Sampling Techniques, is to bias the simulation to visit the conformational space of interest. There are dozens of sampling techniques (see reference (*76*) for an excellent review), here I briefly present the most used ones. Also, I will extend on Maxwell-Demon MD and Metadynamics due to their relevance to this Thesis.

- **Targeted MD** (*199*)

  It enforces the transition between two known states of the system (A and B) introducing energetics restrains to the trajectory. The restrains gradually reduce the RMSD to the target state. The procedure accepts iterations to refine the transition path.

- **Steered MD** (*200-202*)

  Alternatively to Targeted MD but on the same principle, a pulling spring term is added in the Hamiltonian. The spring acts as steering force towards a known target state, ligand (un)binding or a fixed point in the space. Steered MD is used in processes where external forces can help such as protein unfolding, ligand unbinding or *in sillico* atomic force microscope experiments. Steered MD combined with the work of Jarzynski enables to obtain free energy profiles from non equilibrium simulations (*203*). Several pulling experiments are needed to obtain energy profiles. Extensions of the method allows to capture kinetics rates (*108*, *109*).

- **Accelerated MD** (*204*, *205*)

  Accelerated Molecular Dynamics modifies the potential energy of the system in a way that reduces the height of energetic barriers, mainly local ones. It was introduced by the McCammon group to simulate activated processes. The method requires only one copy of the system, reaching significantly better sampling over the conformational space than traditional MD (*206*). With proper Boltzmann re-weighting the original free energy profile is recovered.

- **Replica Exchange Simulations** (*207*, *208*)

  Replica Exchange simulations uses multiple trajectories of the same system that are periodically interchanged. The simulation ensemble spans a range of meaningful temperatures by setting a different temperature to each replica. After a fixed time period individual trajectories are evaluated in a Metropolis test and trajectories are interchanged between temperature levels. In other words, if a high temperature replica meets a low potential energy conformation the replica temperature will smoothly decrease. The replica exchange method was extended to interchange the Hamiltonian instead (or in addition) of the temperature. The main advantage of replica exchange simulations is that

parallelization is trivial, ideal for large CPU clusters. On the disadvantage side, most of the kinetic information of the system is lost.

- **Umbrella Sampling** (*209*, *210*)

  Likely to be the oldest biasing technique was and still is widely used to simulate processes where an initial path can be estimated. By defining a reaction coordinate that drives the transition from an initial to a target state, it forces the system to move in small windows of the path. An extra term is added in the Hamiltonian (commonly harmonic) to guarantee sampling in the region of interest. The bias to the free energy caused by the added potential can be eliminated with statistical techniques (*211*, *212*).

- **Transition Path Sampling**(*213*, *214*)

  In transition path sampling a first guess of a potential pathway is used to start a Monte Carlo procedure aimed to iteratively refine the path. Optimized paths are selected by assessing how efficient the initial state is liked to the target. Transition path sampling outputs an ensemble of path, being one of the most (among biased simulations) accurate methods available for detecting paths.

- **Milestoning Simulations**(*215*) (*216-219*)

  Ron Elber's Milestoning simulation technique follows protein motion through a discrete number of states, obtained from an estimated path. Interestingly, Milestoning allows computing kinetics rates of the conformational transition. Assuming that the flow through the states is stationary (a new simulation is launched at the beginning of the path each time any simulation reaches the target state) leads to the probability of each state in the path and ultimately, to free energy values.

- **Maxwell-Demon MD**(*220*, *221*)

  This technique has many variants that have received different names, among them Maxwell-Demon dynamics and dynamic importance sampling (DIMS). The main idea of the technique is to bias trajectories by introducing information that help them to sample the transition path. It does not perturb the energy landscape but selects slices of trajectories that move towards the target state by using a ratchet-like Metropolis acceptance filter. If $\Delta\varphi$ is an observable that captures the motion of slice of trajectory generated (originally 1 picosecond of MD trajectory), the probability of accepting it (p_{acc}) is defined as:

$$\text{(Eq. 10)}\quad p_{acc}(\Delta\varphi) = \begin{cases} 1 & if\ \Delta\varphi \leq 0 \\ e^{-|\gamma\,\Delta\varphi|^2} & if\ \Delta\varphi > 0 \end{cases}$$

  where γ is a parameter tuned to control the softness of the acceptance rate, typically defined by the initial distance between two known structures. Here, $\Delta\varphi < 0$ means that the proposed slice is moving towards the target structure. Backwards steps, $\Delta\varphi < 0$, can

be accepted yielding not necessarily linear trajectories (*45*, *46*, *222*). If the slice is not accepted, a new one is generated. Repeatedly, the algorithm guarantees net motion towards the target structure, without modifying the system Hamiltonian.

- **Metadynamics**(*223*, *224*)

  Metadynamics eases trajectories to new regions of the conformational space by filling the visited potential energy surface minima with Gaussian function. The algorithm assumes that the system can be described by few collective variables λ, fully locating the trajectory in the conformational space by means of them. Each time that similar conformations are sampled (in λ space) a positive Gaussian function is added to the original Hamiltonian (Eqs. 11,12), discouraging the system to come back to this point. More Gaussians sum up with the evolution of the simulation until the energy landscape is full, at that point, the real free energy landscape is the opposite of the sum of all Gaussians.

$$\text{(Eq. 11)} \quad V(\lambda, t) = \int_0^t dt' \, w \, \exp\left(-\sum_i^d \frac{(\lambda_i - \lambda_i(t))^2}{2\sigma^2}\right)$$

$$\text{(Eq. 12)} \quad H_{total} = H_{initial} + V(\lambda, t)$$

  where λ are the collective variables, t is the time, d the number of collective variables, w and σ are the height and width of the Gaussian added at time. Metadynamics ensure trajectories to escape the initial energy basin being specifically suitable to describe motions beyond equilibrium (*225*), like conformational transitions. Since it is not trivial to identify the stopping point of a simulation, this in principle exact method, suffers from convergence problems. If the simulation is overextended artifactual Gaussians functions will be added, distorting the energetic description. To overcome this problem the well-tempered Metadynamics was introduced where the height of the Gaussian is history-dependent, being smaller as conformations are re-visited (*224*, *226*). Well-tempered Metadynamics is a widely used method with improved convergence in free energy predictions.

After 40 years of the foundation of computational biology, it is clear that to simulate real biological processes, a shift in the paradigm will be needed. Despite its power, MD alone is not going to be able to dissect macromolecular complexity. New innovative approaches need to be developed.

Chapter 2: Objectives

This Thesis is about methodological developments to trace conformational transitions in proteins. Here I present the problem description and the open issues where efforts were dedicated.

2.1 Understanding Conformational Transitions

Proteins and macromolecules in general constantly move from one conformation to another one at physiological conditions (*76*). Motions range continuously from small local rearrangements, to large domain translocations including unfolding/folding events. Dissecting protein motions and more importantly their causes is key to understand protein function (*227-229*). One example in the way to rationalize protein conformational changes has been the traditional division between induced-fit and conformational selection binding modes. The broadest objective of this Thesis is to contribute to the understanding of conformational motions, in this regard we want to develop physical models that reduce the complexity of biological systems to investigate those movements.

2.2 Simulate Protein Motions

Molecular simulations play a central role in understanding protein dynamics. Particularly, the identification of the transition path between conformations is an open field of research that disclosed several successful strategies. The obvious one is pure force MD simulations, that when the system and computational power allows, gives the best results (*230*). Enhanced sampling algorithms based on atomistic MD have been extensively used to simulate conformational transitions but the problem is far from being solved. Enhanced sampled methods require high user expertise, are CPU demanding and they have a common drawback: they require two known conformers, an estimated transition path, or alternatively the collective variables describing the transition. One objective of this Thesis is to design strategies to simulate conformational transitions in proteins. The aim is to use a multiscale approach combined with algorithm heuristics to reduce the computational cost of simulations. One major aspect to solve is to include at each stage the adequate level of resolution ensuring that the study is feasible.

2.3 Exploit Simple Models

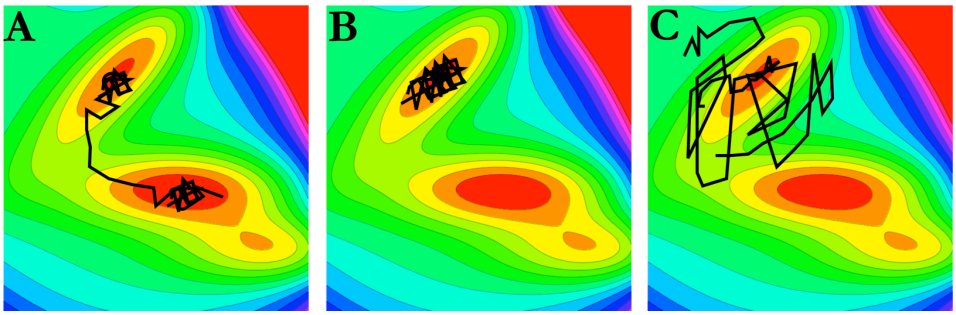
On top of the theoretical interest for better algorithms, simpler models are easier to understand and give valuable information where traditional methods fail (*62*, *112*, *134*). At initial exploratory level, solvent effects will be accounted for via implicit solvation potential to gain computational efficiency. For this purpose we have adapted the effective energy function EEF1 proposed by

Lazaridis and Karplus (*92*). The next item is to find the balance in the protein representation. Aware of the fact that representation may need to adapt for specific purpose, proteins models will be constructed with a variety of resolutions, from atomistic to a limited number of particles per residue in the investigatory phase of the project. Another aim of this Thesis is to assess the suitability of CG models for the exploration of conformational transitions in combination with main energetic descriptions. Ideally, protein models with predictive power will be developed and tested with applications in biomedicine.

2.4 Develop Efficient Computational Tools

Although a considerable improvement has been achieved during last decade, functional protein dynamics happen on a timescale out of the scope for the most accurate computational methods. As discussed, the computational burden of MD prompted alternatives methods to trace conformational transitions. Figure 6 shows some examples of algorithms used to boost MD efficiency. We want to combine these techniques with discrete Molecular Dynamics sampling engine. To this end, we want to design reliable tools to trace conformational transitions in proteins when both ends of the transition path are known.

Figure 6: Shortcomings of tracing transitions paths



Conformational transitions modeling problems. (A) Example of desirable transition path: dark lines shows a simulation trajectory between the starting structure (red basin, up) and target structure (red basin, down). (B) Usually, simulations are stuck in the initial energy basin, impeding correct sampling. (C) Once the initial basin is abandoned there is no guarantee to visit the target basin due to the complexity and ruggedness of proteins energy landscape.

2.5 Discrete Molecular Dynamics for Conformational Transitions

The lesson learned with the success of the duo ENM-NMA on describing conformational transitions points in two directions. First, basic sample schemes can identify the direction of simple conformational transitions. The problem arises when larger topological changes are required. In this regard, dMD simulations met the compromise between efficiency and sampling quality. From protein folding we know that dMD can reveal folding pathways (*58*, *63*, *64*, *231*),

so the topological change is covered. Moreover, dMD proved its ability to reproduce the protein equilibrium dynamics (*112*, *232*). The second point is that conformational can be captured with very simple potentials, or at least the main features.

The main goal of this Thesis is to explore dMD capabilities to study drastic movements in proteins. The project will require design specific-purpose methodologies, implement both Structure-Based and Physics-Based force field at different stages and investigate its relative efficiency. Methodological improvements as well as algorithms to sample the relevant conformational space are goals of this Thesis, including re-parameterization and refinement of the associated force field.

2.6 Predicting Protein Conformers

To estimate transition path when only one conformation is known, we need a way of identifying the target state, ultimately predicting its structure. It is clear that while waiting for faster sampling methods and better force field the single way to reproduce an unknown conformational transition is to plug in orthogonal information into the Hamiltonian. The nature of this information is fairly irrelevant as long as it captures the dynamical traits of the protein. For instance, De Groot and co-workers provided an intuitive example where they used a known conformation and the radius of gyration (Rg) of the *unknown* conformer to bias the trajectory to sample configurations that match that Rg value (*233*). Combining small angle X-ray scattering (SAXS) and CG simulations Hummer and co-workers followed a conformational transition of the complex CHMP3 of ESCRT-III, a protein with multiple helical domains separated by flexible linkers (*234*). Also, using SAXS technique, Hub and co-workers biased trajectories to reproduce experimental measurements yielding a spontaneous conformational transition for ribose-binding protein (*235*). In a pure *in silico* approach Onuchic and co-workers used sequence information to predict alternative conformers in proteins (*236*). In this line, we will adapt dMD formalism to interplay with experimental and bioinformatics data, leading to algorithms to predict proteins conformers and the corresponding transition path when only one structure is known.

2.7 Make tools available

We want to make user-friendly tools that translate the methodological advances achieved in this Thesis, ideally building a framework that facilitate the simulation of conformational transitions.

Chapter 3: Atomistic transition path from dMD simulations

In this work we present a new method to trace conformational transitions in proteins in atomistic resolution. There are two types of approaches in modelling conformational transitions, physical simulations and non-physical morphing between two known structures. Morphing methods can be used as an estimator of the transition path, as hypothesis generator or simply to visualize protein motions (*237-240*). They major advantage is that they can complete the conformational transitions in minutes, which is very convenient at exploratory levels. But they suffer from shortcoming due to their simplistic nature. Morphing methods can violate bond distance or obtain surrealistic transition paths, like between a structure and its mirror image. Even sophisticated morphing methods that take into account physical properties like energy display abrupt jumps in potential energy (*238*), unlikely to occur in the cell.

Our protocol, named Maxwell-Demon dMD (MDdMD), is part of a third generation of methods (*238*, *241*, *242*) that are computationally as efficient as morphing methods without loosing the physical nature. MDdMD can complete a conformational transition in about 2 hours of computational time, with atomistic resolution. Trajectories are biased to approach the target structure in a very soft fashion, where the analogy with Maxwell-Demon comes from selecting slices of trajectories that satisfy a spatial restriction. MDdMD paths are energetically smooth, obtaining by sampling structures with a physical force field that guarantees chemically correct structures. We validated our results with known experimental intermediates that our protocol spontaneously sampled in all five cases. Sampled conformers can be used as starting point of MD simulations with minimal adaptation obtained from automated methods.

This study shows an example of the divide-and-conquer strategy to sample conformational space in proteins. After an ensemble of conformers is obtained from fast dMD simulations, we could use each of them as starting points of MD simulations. This can partially overcome the sampling problem since it avoids kinetics traps of the initial structure.

Title: **Finding Conformational Transition Pathways from Discrete Molecular Dynamics Simulations**

Authors: Pedro Sfriso, Agustí Emperador, Laura Orellana, Adam Hospital, Josep Lluis Gelpí, and Modesto Orozco

Stage: Published

Journal: Journal of Chemical Theory and Computation

Type: Research Article

Supplementary Material: http://pubs.acs.org/doi/suppl/10.1021/ct300494q

Author Contribution: PS was the main responsible all the work, developed the method and ran the simulations. PS contributed to the writing the paper.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Finding Conformational Transition Pathways from Discrete Molecular Dynamics Simulations

Pedro Sfriso,[†] Agusti Emperador,[†] Laura Orellana,[†] Adam Hospital,[†,‡] Josep Lluis Gelpí,[†,§,∥] and Modesto Orozco*[,†,‡,∥]

[†]Joint IRB-BSC Program in Computational Biology, Institute of Research in Biomedicine, Josep Samitier 1-5, Barcelona, 08028, Spain

[‡]Structural Bioinformatics Node, Instituto Nacional De Bioinformática, Institute of Research in Biomedicine, Josep Samitier 1-5, Barcelona, 08028, Spain

[§]Computational Bioinformatics Node, Instituto Nacional De Bioinformática, Barcelona Supercomputing Center, Jordi Girona 29, Barcelona, 08034, Spain

[∥]Departament de Bioquímica, Facultat de Biologia, Universitat de Barcelona, Avgda Diagonal 647, Barcelona, 08028, Spain.

Supporting Information

ABSTRACT: We present a new method for estimating pathways for conformational transitions in macromolecules from the use of discrete molecular dynamics and biasing techniques based on a combination of essential dynamics and Maxwell−Demon sampling techniques. The method can work with high efficiency at different levels of resolution, including the atomistic one, and can help to define initial pathways for further exploration by means of more accurate atomistic molecular dynamics simulations. The method is implemented in a freely available Web-based application accessible at http://mmb.irbbarcelona.org/MDdMD.

INTRODUCTION

Proteins are dynamic entities, whose conformations change in response to a variety of external stimuli, such as temperature, solvent composition, presence of ligands, and electric or mechanical fields.[1] There is now an overwhelming amount of evidence showing that protein function is directly related to protein flexibility,[2−4] and it is clear that evolution has made an effort to not only optimize protein structure but also to design flexibility patterns optimal for function.[5−10] Furthermore, it seems that evolution has used very often the intrinsic flexibility patterns of ancestor proteins to create new macromolecules,[2,11] in a conservative strategy to maintain fold and flexibility. Databases are full of examples where the same protein is found in different conformations due to the presence (or absence) of different ligands.[8,12,13] However, there are very few examples of experimental characterization of protein conformational transitions, since dynamic high-resolution techniques are still in their infancy,[10,14−16] which implies that most of the knowledge from conformational transitions in proteins is derived from molecular simulation techniques.[1,17−34]

Atomistic molecular dynamics (MD) is probably the most accurate and universal simulation technique for the study of protein flexibility. MD is based on a rigorous theoretical formalism and uses physical potentials (the force field) that have been refined and optimized for decades.[35,36] Unfortunately, practical use of MD is limited by the gap between the transition and the currently accessible simulation times, which precludes the use of direct-unbiased MD approaches to study large conformational changes. As a response to this problem, a variety of techniques have been developed to force the sampling along the direction of the transition.[37,38] These biasing techniques can provide encouraging results[38−45] but are very expensive computationally and can lead to incorrect results if the transition coordinate or the restraint protocols are not well tuned. In this complex scenario, morphing coarse-grained (CG) models have been gaining importance as an inexpensive alternative to obtain first guesses of the transition paths.[5−9,13,21,25,46−63]

Within the morphing CG paradigm, the protein is represented at the C_α level ignoring side chains. Transitions are simulated using different approaches, the simplest ones are based on interpolation schemes between original and final conformations using either Cartesian or internal coordinates.[64−69] A more physical variant of the morphing CG method is based on the assumption that evolution has created proteins precoded to perform biologically relevant transitions, which means that biologically relevant conformational transitions are likely to happen along soft modes,[13,53,70−73] i.e., the easiest deformation modes of proteins. These morphing methods require a definition of the protein Hamiltonian, which is often obtained by means of the elastic network models (ENM), where the *minimal frustration* principle[74] is assumed, and accordingly the perturbation energy associated with deformations of protein conformations from known experimental structure follow a harmonic behavior:

$$E = \sum_{i,j} \delta_{ij} K_{ij} (R_{ij} - R_{ij}^0)^2 \quad (1)$$

where i and j are residues, δ_{ij} is a delta function equal to 1 when i and j are at less than a given distance and 0 otherwise, K is a spring constant (linear or distance dependent), R_{ij} stands for inter-residue distance, and the superscript 0 refers to the value of R_{ij} in the reference structure.

Received: June 14, 2012
Published: August 23, 2012

The preferred deformation modes, i.e., those along which the energetic cost of deforming a protein is minimum, are obtained by diagonalization of the Hessian matrix associated with the Hamiltonian outlined in eq 1 (normal-mode analysis; NMA). Methods based on the EN-NMA approach provided then a guess of the transition by activating the movements along low frequency modes overlapping with the transition vector. Transitions obtained by animating natural deformation modes of proteins are more realistic than those obtained by simple interpolation schemes.[70] Unfortunately, EN-NMA morphing approaches also present some shortcomings; a very clear one is the corruption in the covalent structure of the protein related to large displacements along a limited number of modes. To partially alleviate these problems, different authors[51,70] have recalled the principle of minimal frustration also along the transition, and accordingly, lower modes are recomputed for intermediate structures obtained along the transition. Other authors[52,53] have developed EN-NMA in the internal coordinate space. Unfortunately, none of these elegant approaches is useful when transition is not coded in the first essential deformation modes of the protein, or when it implies side chain movements ignored in a C_α representation.

In this paper, we present a new method to trace large conformational transitions based on a very fast discrete molecular dynamics (dMD) algorithm. Plausible trajectories are obtained by following ballistic equations of motions are biased toward the target structure by means of a Maxwell–Demon engine (see below), which incorporates information of essential deformation movements of the protein. The method can work with any level of resolution (including all-atoms and hybrid levels), is very efficient computationally, and displays very good performance in a large variety of test systems.

METHODOLOGICAL APPROACH

Basic Discrete Molecular Dynamics Algorithm. The basic dMD formalism[62,75−77] assumes that particles move in the ballistic regime (constant velocity) until a particle−particle collision occurs. In dMD, the potential energy is defined with stepwise discontinuous functions of the particle−particle distance instead of continuous functions used in standard molecular dynamics. In the absence of any collision, the particles move linearly with constant velocity. The position of a given particle at the time of the next collision is

$$\vec{r}_i(t + t_c) = \vec{r}_i(t) + \vec{v}_i t_c \tag{2}$$

where $\vec{r}_i$ and $\vec{v}_i$ stand for positions and velocities and t_c is the minimum among the collision times t_{ij} between each pair of particles i and j:

$$t_{ij} = \frac{-b_{ij} \pm \sqrt{b_{ij}^2 - v_{ij}^2(r_{ij}^2 - d^2)}}{v_{ij}^2} \tag{3}$$

where r_{ij} is the square modulus of $\vec{r}_{ij} = \vec{r}_j - \vec{r}_i$, v_{ij} is the square modulus of $\vec{v}_{ij} = \vec{v}_j - \vec{v}_i$, $b_{ij} = \vec{r}_{ij}\cdot\vec{v}_{ij}$, and d is the distance corresponding to the wall of the square well.

When two particles collide, there is an elastic transfer of linear momentum into the direction of the vector $\vec{r}_{ij}$:

$$m_i\vec{v}_i = m_i\vec{v}_i' + \Delta\vec{p} \tag{4}$$

where the prime denotes the variables after the collision. The new velocities after collision are obtained by applying conservation rules:

$$m_i v_i + m_j v_j = m_i v_i' + m_j v_j' \tag{5}$$

$$\frac{1}{2}m_i v_i^2 + \frac{1}{2}m_j v_j^2 = \frac{1}{2}m_i v_i'^2 + \frac{1}{2}m_j v_j'^2 + \Delta V \tag{6}$$

where ΔV stands for the depth of the square well defining the interatomic potential.

The transferred momentum can be easily determined from

$$\Delta p = \frac{m_i m_j}{m_i + m_j}\left\{\sqrt{(v_j - v_i)^2 - 2\frac{m_i + m_j}{m_i m_j}\Delta V} - (v_j - v_i)\right\} \tag{7}$$

Note that the two particles can go out of the well as long as

$$\Delta V < \frac{m_1 m_2}{2(m_1 + m_2)}(v_j - v_i)^2 \tag{8}$$

Otherwise, if the particles remain in the well, eq 7 reduces to

$$\Delta p = \frac{m_i m_j}{m_i + m_j}\{\sqrt{(v_j - v_i)^2} - (v_j - v_i)\} \tag{9}$$

which, taking the negative solution of the root, leads to

$$\Delta p = \frac{2m_i m_j}{m_i + m_j}(v_i - v_j) \tag{10}$$

In summary, in dMD no forces should then be calculated; the equations of motion should be integrated on the femtosecond scale, yielding then to very efficient simulations.[58,60,78−80] Previous works in our group have shown how dMD is able to reproduce well equilibrium dynamics of proteins as determined by explicit-solvent atomistic MD simulations[58,60,79] and can be used to perform robust minimizations of protein−protein complexes (manuscript in preparation). Other authors demonstrated the sampling capability of dMD folding small proteins.[76,81] Also, encouraging results from dMD have been reported in the analysis of many aspects of protein and nucleic acids dynamics,[61−63,77,82−85] macromolecular aggregation,[56,86−88] and macro- and supramolecular transitions.[89]

Force Field Description. dMD is based on the use of simple or multiple-step square potentials to describe physicochemical interactions. Our dMD interaction potentials include an implicit solvation term (derived from Lazaridis−Karplus functions[90]) and van der Waals and electrostatic terms. In the aim of increasing the speed of the simulations, we have chosen our simplest version of dMD:[59] in the case of attractive interactions, two-step potentials that define square wells are used, and a soft barrier in the case of repulsive interactions. As in standard dMD calculations, infinite wells were used to maintain all bonds and angles in the protein near equilibrium values, preventing then distortions of the chemical backbone. In this implementation of the method, well-defined secondary structure elements were enforced during the transition by defining very deep square potentials between hydrogen bonded groups. These secondary-structure constraints are automatically released in cases where initial and final secondary structures do not match.

Biasing Techniques. The core of our morphing procedure is a biasing algorithm, which enhances dMD sampling in the direction of the transition and also, if possible, along the essential deformation modes (Figure 1). The first is quantified by simple metrics, which in principle are applied only to the C_α's:

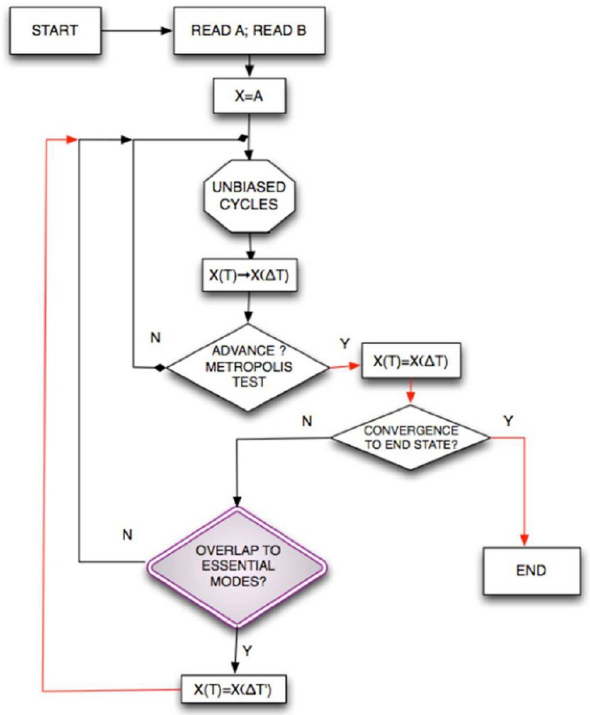


Figure 1. Flowchart of the basic MDdMD algorithm. Red lines indicate a definitive temporal advance toward the target structure. NMA biasing criteria (purple) only apply under certain conditions (see Figure 2).

$$\Gamma = \sum_{i=1}^{N} \| \vec{r}_{i,\mathrm{B}} - \vec{r}_{i,\mathrm{X}} \| \omega(i) \tag{11}$$

where N is the total number of residues, B is the target structure, X is the sampled conformation (X = A for the original conformation), and $\omega(i)$ is a weighting function defined as

$$\omega(i) = \begin{cases} 0 & \text{if } \| \vec{r}_{i,\mathrm{B}} - \vec{r}_{i,\mathrm{X}} \| < r_{\mathrm{cut}} \\ \| \vec{r}_{i,\mathrm{B}} - \vec{r}_{i,\mathrm{X}} \| & \text{if } \| \vec{r}_{i,\mathrm{B}} - \vec{r}_{i,\mathrm{X}} \| > r_{\mathrm{cut}} \end{cases} \tag{12}$$

where r_{cut} is an estimate of oscillation around equilibrium structures generated by the thermal noise (1.5 Å from the study of our MODEL database[91]).

We define the transition vector ($\Delta\vec{R}$) as that connecting the sampled structure to the target one ($\vec{R}_B - \vec{R}_X$) and the essential transition vector defined from the combination of eigenvectors that better reconstructs the transition:

$$\vec{V}_{\mathrm{NM}} = \sum_{j=1}^{m} \alpha_j \vec{v}_j \tag{13}$$

where $\vec{v}j$ stands for the eigenvectors obtained from normal-mode analysis (see below), and α_j is the normalized overlap between the selected mode (j) and the transition vector. In order to reduce the noise, the sum extends for the m modes with $\alpha_j > 0.15$ pertaining to the group of 10 lower frequency ones (the essential deformation space).

Following a Maxwell−Demon approach, the bias toward the target structure is not introduced by an energy penalty, but using informational criteria.[39,92] Thus, after at a certain simulation step (t), the progress variable (Γ) is computed (eq 11) and compared with that obtained in a previous accepted movement ($t - \Delta t$). Following the Metropolis Monte Carlo procedure, the simulation segment ($t - \Delta t$)→(t) is preaccepted or not based on the probability p_t:

$$p_t = \begin{cases} 1 & \text{if } \Gamma_t < \Gamma_{t-1} \\ \exp\left[-\frac{1}{2} \left(\frac{\Gamma_t - \Gamma_{t-1}}{\beta \mathrm{RMSD}_{(X,t,B)}} \right)^2 \right] & \text{if } \Gamma_t > \Gamma_{t-1} \end{cases} \tag{14}$$

where β is dynamically adjusted to guarantee a user-input acceptation rate (recommended value around 40%), and the time frame (Δt) is typically 100 RTU (see eq 15; in our experience, 1 RTU corresponds to around 10−20 ps of standard protein dynamics).

$$\mathrm{RTU} = \frac{0.15 \text{ Total Collisions}}{\#\text{residues}} \text{ at } T = 300 \text{ K} \tag{15}$$

The protocol outlined above is very efficient for driving transitions (without aberrant contacts or distortions of chemical structure), but it does not guarantee that such a transition follows essential deformation movements, which in some cases might bias the transition to biologically unlikely paths. To guarantee that, if possible, transition uses the default conformational flexibility of the protein, we compute the overlap between the essential transition vector ($\vec{V}_{\mathrm{NM}}$) and the transition vector ($\Delta\vec{R}$), taking first $\vec{V}_{\mathrm{NM}}$ defined from the eigenvectors of the original structure (A). We found that, if the overlap is above 0.6, we can consider that essential deformation space contains information on the transition and proceed as described in the following paragraph. When the overlap is smaller than this cutoff, we consider that the essential deformation space does not contain useful information to improve our definition of the transition pathway, and transition is fully guided by the dMD force field and the Maxwell−Demon.

Assuming that there is a good overlap between the transition vector and the essential transition vector defined from eigenvectors of the original structure (X = A), we incorporate additional conditions into the preaccepted configuration as defined by probability function:

$$p_{\mathrm{NM}} = \begin{cases} 1 & \text{if } \tau > T \\ \exp\left[-\frac{(T - \tau)^2}{\beta_{NM}} \right] & \text{if } \tau < T \end{cases} \tag{16}$$

where the index τ is the normalized projection between two structures separated by a significant period of time ($\Delta t'$) in the range 20−100 RTU ($\Delta t'$ is automatically adjusted depending on protein size) and T is by default 0.6. The β_{NM} factor is set to 0.1 to guarantee a smooth evolution of the probability function in the range 0.15−0.6.

As noted by others (see the Introduction), when the structure moves from the starting conformation (i.e., X ≠ A), the principle of minimal frustration is not granted, and accordingly eigenvectors computed for A lose predictive power. Thus, when our algorithm detects that the transition vector between target structure and current transition structure at $t + \Delta t'$ ($X_{t+\Delta t'}$) does not overlap with the essential transition vector determined from the eigenvectors for the starting structure (X = A), it assumes that original essential deformation space is no longer informative. At this point, if the transition structure ($X_{t+\Delta t'}$) is close to the final conformation (X = B; based on a spherical cutoff

dx.doi.org/10.1021/ct300494q | J. Chem. Theory Comput. 2012, 8, 4707−4718

adjusted to the size of the protein, typically around 3.5 Å), we use the eigenvectors of B to compute the essential transition vector. Otherwise, we recomputed the eigenvectors assuming that structure $X_{t+\Delta t'}$ is a minimum and applied the same protocol outlined before (see Figures 1 and 2).

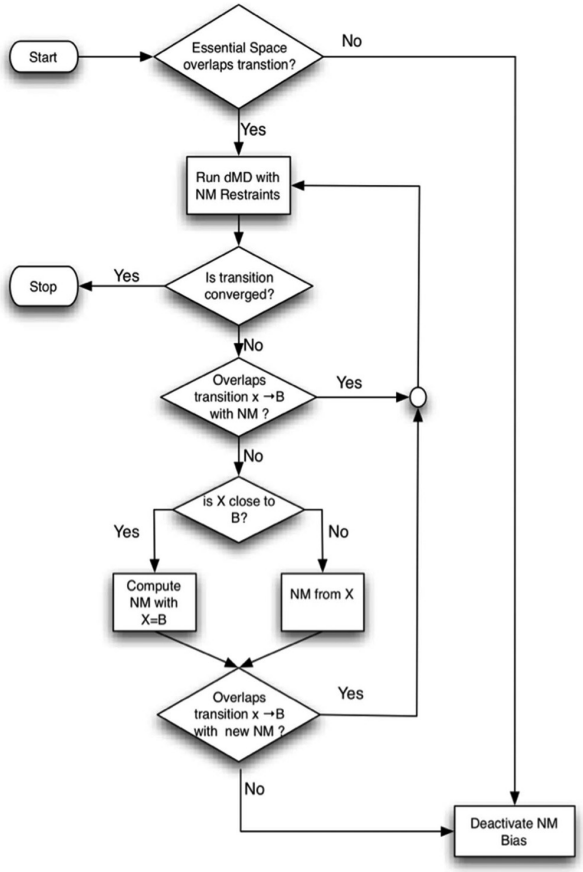


Figure 2. Detail on the implementation of the NMA bias based on the initial and current overlap between transition and essential deformation space.

In a small number of cases, especially when transition is close to the target structure, the essential deformation movements are not very useful for driving the transition. In this case, eq 16 increases dramatically the rejection rate, making final convergence slower. The algorithm recognizes automatically this situation, inactivating from this point of the introduction of information on essential modes as a guide for the transition (Figure 2).

Computing Normal Modes. The essential deformation space of a given structure X is computed using our version of the elastic network model. It is based on a Go-like harmonic potential with differential sequential and spatial cutoff functions and a Kovac's distance dependent force constant adjusted to reproduce essential dynamics obtained from atomistic molecular dynamics simulations (see refs 13 and 80 for details of the method). It is worth noting that there is a very good agreement between the type of flexibility predicted by this EN-NMA model and that obtained by dMD simulations,[59,60] which guarantees the physical consistency of the present hybrid model. The EN-NMA routine is incorporated in our code for recomputing eigenvectors when required.

Convergence into Target Basin. It is not trivial to determine when a transition has reached the target structure, since protein structures are continuously moving by thermal noise (in average 1−2 Å based on atomistic MD simulations). Bearing this in mind, we have decided not to attempt to reach a very low RMSD to the target conformation. In fact, even for a perfect force field, it is unrealistic to obtain a zero RMSD structure when a protein structure is naturally oscillating at room temperature.[22]

Thus, we adopt here a convergence criterion based on the slope of RMSD to target structure respect to time. This criterion prevents our method from obtaining stressed structures that would be only an artificial minimum as a result of the strong bias necessary to achieve close to zero RMSD values, but the user should be aware that a further refinement using explicit solvent atomistic MD simulation might be necessary. Thus, in our procedure, below a user-provided RMSD cutoff (recommended values around 2 Å for a 200 residues protein to 3 Å for a larger proteins around 600−700 residues), we conclude that the system has reached the target equilibrium state (near target experimental structure) when in the last 20−40 Reduced Simulated Time Units (RTU; see eq 15) of dynamics there is a negligible change in RMSD.

RESULTS AND DISCUSSION

Transition Characterization. We tested the ability of our method to obtain reasonable estimates of pathways for conformational transitions by exploring a large database of cases where there are at least two clearly different structures in the protein data bank (PDB[93]). The validation data set (see Table 1) contains 47 protein pairs ranging from small (around 100) to very large (around 1000) proteins. In all cases, trajectories have been followed in both directions, yielding a total of 94 transitions, all of them followed at the all-heavy-atoms level of resolution. Conformational changes in the flexible part (obtained by computing RMSD after manual alignment of rigid part) vary from small (less than 3 Å) to very large (more than 20 Å), which means that we are trying to reproduce not only trivially small transitions but also massive conformational changes (see Figure 3), which are more challenging for transition pathway detectors. Analysis of the overlap between the transition vector and the lowest eigenvectors of the equilibrium conformations shows in general a reasonable overlap between the transition and the essential deformation pattern of proteins (see Table 1 and Figure 3), confirming that often biologically relevant conformational changes follow the essential deformation modes of proteins.[3,69] However, 68 of the 94 transitions display overlaps below 75%, and 34 of them below 50% (Figure 3), meaning that a significant number of transitions cannot be simply explained by the essential deformation pattern of proteins. Finally, it is worth noting that the reversibility in the transitions is not always granted, since overlaps between essential deformation and transition vectors are significantly different when considered in the A→B and B→A directions for a non-negligible number of cases (see Table 1).

A detailed analysis of the entire data set reveals that 44% of transitions correspond to conformational changes between the open and closed forms of the protein, 39% to induced binding to other macromolecules, 60% to the binding of small ligands or cofactors, and one requires post-translational modifications (see Table S1). The database contains no example of trivial

Table 1. List of Proteins Considered in the Validation of the Method[a]

structure pair	# residues	RMSD (Å)	overlap	ΔRg/Å	difficulty
1yzy*/2ahm*	71	7	0.39/0.45	0.16	++
1szv*/1vet*	91	5.24	0.23/0.26	1.05	++
1l5e/1l5b	101	6.7	0.81/0.83	0.31	−−
1wrp/3wrp	108	2.48	0.49/0.46	0.61	+−
1xfr*/2fjy*	123	5.84	0.30/0.32	0.95	++
1e7xA/1dzsB	129	3.4	0.07/0.10	0.90	+−
1cfd/1cfc	148	5.43	0.94/0.92	1.74	−−
1h2d*/1oc3*	158	8.6	0.27/0.27	1.53	++
2gja/1rfl	162	8.71	0.20/0.33	2.04	+−
1r3e*/1ze1*	169	5.94	0.29/0.33	0.22	++
1ybj*/1dk0*	173	5.64	0.24/0.22	0.06	++
1aje*/1ees*	174	6.82	0.39/0.56	1.81	++
1cbuB/1c9kB	180	3.55	0.38/0.52	0.32	+−
1ex6/1ex7	186	3.64	0.88/0.79	0.83	−−
1s2h*/1go4*	190	4.9	0.48/0.56	0.28	+−
1bcc/2bcc	196	7.45	0.53/0.54	0.52	+−
2rh5/2rgx	202	5.99	0.88/0.62	2.51	+−
4ake/1ake	214	7.19	0.88/0.62	3.08	−−
1ggaA/1wdnA	220	5.34	0.86/0.61	1.56	−−
2lao/1lst	238	4.81	0.90/0.59	1.53	−−
3pjr*/1qhh*	261	9.38	0.47/0.61	0.33	+−
1urp/2dri	271	4.24	0.92/0.88	1.01	−−
1ram/1leiA	273	3.38	0.94/0.90	1.46	−−
5at1/8atc	310	2.59	0.66/0.45	0.50	+−
1 cmkA/1 cmkB	317	3.62	0.94/0.72	1.27	−−
3dap/1dap	320	4.35	0.90/0.74	1.30	−−
1eyk/1nuz	327	4.54	0.56/0.28	1.00	+−
1bp5/1a8e	329	6.81	0.86/0.67	1.98	−−
1jqj/2pol	366	1.99	0.48/0.38	0.46	+−
1omp/1anf	370	3.91	0.86/0.86	0.90	−−
8adh/6adh	374	1.27	0.24/0.31	0.16	+−
9aat/1ama	401	1.67	0.35/0.44	0.27	+−
1ux5/1y64	411	10.33	0.84/0.61	3.75	++
1qf5/1hoo	431	3.03	0.32/0.56	0.65	+−
1yyo/1yyw	438	17.96	0.14/0.40	2.37	++
1bnc/1dv2	452	4.51	0.83/0.79	1.56	−−
1rkm/2rkm	517	3.24	0.92/0.66	0.58	−−
1sx4/1oel	524	12.61	0.77/0.76	3.87	−−
1hp1/1hpu	525	9.93	0.52/0.53	0.34	++
2hmi/3hvt	556	3.45	0.59/0.60	0.61	+−
1i7d/1d6m	620	3.65	0.61/0.48	0.08	+−
8ohm/1cu1	645	4.62	0.77/0.71	0.29	+−
1lfg/1lfh	691	6.54	0.76/0.85	1.08	−−
1qvi/1kk8	837	27.61	0.72/0.78	1.70	+−
1q9x/1q9y	899	5.91	0.81/0.46	0.57	+−
1ih7/1ig9	903	6.47	0.77/0.48	0.68	+−
1su4/1iwo	994	13.93	0.71/0.56	0.92	+−

[a]For each transition, we quoted the starting and final structures (in PDB code), the size (in number of residues) of the protein, the root mean square between the two structures, the change in radius of gyration, the overlap between the transition vector, the expected difficulty of the transition (see main text), and the essential deformation space of the reference proteins (first 10 modes, overlap is computed for each pair (a/b) in both directions: a→b, first number and b→a, second number).

conformational changes, but it is clear that the degree of complexity of finding a reasonable transition path is not homogeneous across the data set. Thus, we classified transition

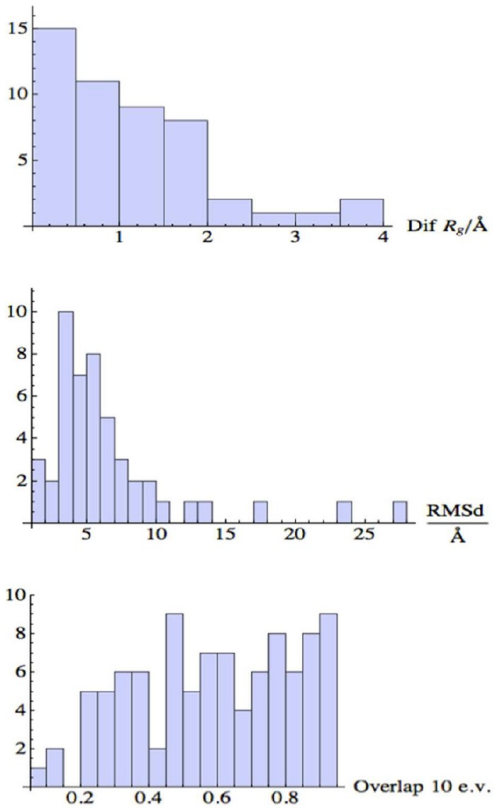


Figure 3. Distribution of difference in radius of gyration (top), between starting and final structures, RMSD between both stats (middle), and overlap of initial essential deformation space with transition (bottom) for all the transitions considered.

between all protein pairs in three categories (simple (−−), difficult (−+), and very difficult (++)) based on three simple descriptors: (i) RMSD between the pairs, (ii) the maximum overlap between the transition vector and essential deformation modes, and (iii) the expected reversibility. On the basis of this, our extended database contains 33% simple transitions, 46% difficult transitions, and 22% very difficult transitions (see Table S1).

Global Performance of the Method. We managed to reach the final structural basin for 91 of the 94 transitions (see Table S2), without any *ad hoc* adjustment of the method for difficult cases. Looking in detail to the final structures, we did not find local errors, like too large bonds or unrealistic angles or dihedrals, and only in three cases did we detect some problems in local geometry (see Table S2) corresponding to cases where experimental transition implies a disruption of elements of the secondary structure, which were not captured by our default setup. In summary, our method has a maximum failure ratio of only 6% when exploring a large database of transitions, some of them very difficult to trace (see Table 1). From 94% of the successful cases, we can detect a few cases where the transition is too small and there is overlap between the starting and final basins, making it difficult to guarantee the success of the transition. In these cases (six cases from the 96 considered; see Table S2), direct atomistic studies based on umbrella sampling, targeted MD, and metadynamics of alternative biasing techniques seem a more sensible election than coarse-grained approaches.

It is worth noting that essential deformation as defined by normal-mode analysis helps to find a reasonable pathway in around 70% of the cases; even in a significant number of cases (36%), the reference structure used to determine the essential deformation space was changed in the course of transition (see Methodological Approach). The method managed to obtain physically plausible transitions (see Figure 4) in cases of dramatic changes (see for example 1qvi←→1kk8 or 1sx4←→1oel transitions in Table S2), in very large systems (see for example 1su4←→1iwo or 1ih7←→1ig9), and also in cases (like 1ybj←→1dk0 and 1h2d←→1oc3) where there is a very poor overlap between essential deformation space and the transition vector (see Figure 4 and Table S2). Using the reduced simulation trajectory time (RTU relative to size and RMSD of the transition), we can evaluate the real difficulty of our procedure to reach the target structure. As shown in Table S3 and Figure S1, the method finds transition paths very quickly in most cases. In general, the cases (Figure S1) where pathways are difficult to find correspond to transitions labeled as "very difficult" in Table 1, typically cases where essential deformation space is quite orthogonal to the transition, and where there is a large ratio of rejection by the Maxwell–Demon algorithm. Translation of reduced simulation time to wall-clock time is difficult, but for most (60%) of the transitions outlined here, the algorithm finishes in less than 1–2 h on a standard laptop computer. Transition pathways obtained with MDdMD are consistent with all the structural parameters considered in our physical based force field, though we expect them to be a valuable initial estimation. However, we should stress that the first guess of transition path may still not be the kinetically optimal one, so further refinements using much more rigorous (and computationally expensive) methods might be needed.

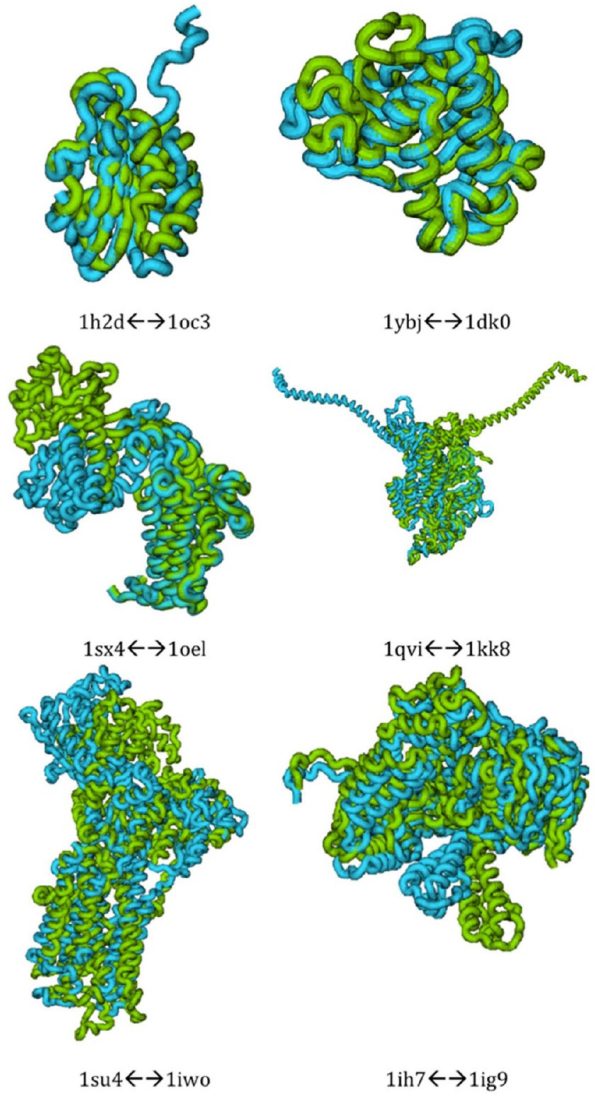


Figure 4. Structural ribbon superposition of both experimental ends of some nontrivial conformational transitions explored.

Reversibility. Experimentally, there are many transitions which are not equally easy in both directions (A→B and B→A). This is clearly visible in cases where the overlap between the transition vector A←→B and the essential deformation spaces of A and B are very different (see Table 1), warning about reversibility problems in our simulations. Fortunately, even the reduced simulation times in the A→B and B→A directions can be quite different for some pairs of proteins (see Table S2); there is only one case (adenosylcobinamide kinase, 1cbu/1c9k) where our method shows irreversibility. Very interestingly, the analysis of the RMSD bidimensional plots (see examples in Figure 5) indicates that for a given point in the transition paths, structures sampled in the A→B and B→A directions are quite similar, suggesting microscopic reversibility in the transition. The same is clear by looking at the evolution Maxwell–Demon acceptance rate and the RMSD (to target) along the transition (see examples in Figure 5). In summary, our method, which is not based on interpolation or on the use of geometrical energy restraints, is able to find with reduced computational effort feasible and (macro- and microscopically) reversible transitions between distant conformations.

Local Geometrical Quality. One of the main problems of interpolation techniques and NMA-based morphing techniques relies on the lack of quality of the local geometry, which might display unrealistic bond lengths or angles, or even steric clashes. Our procedure, which is based on a very simple, but still reliable, physical potential, eliminates completely these problems. No violation of chemical connectivity or steric clashes are found during transitions. PROSA profiles[94] reveal that not only final structures but also generated intermediate conformations fulfill the standard requirements of a folded protein (see examples in Figure 6). Violations of Ramachandran's maps are quite small along the transitions (Figure 7), and the pattern of native contacts predicted by our method matches in general that found experimentally (Figure 7). It is very encouraging that even the current version of the method uses only the deviation from the target structure of the C_α's as a decision variable for the Maxwell–Demon selection procedure; side chains are typically well positioned (see examples in Figure 7). In any case, it is worth it to note that any error in side chain positioning can be easily corrected in our algorithm by just including information-biased torsion parameters in the dMD potential function.

Sampling of Transition Intermediates. The procedure outlined here allows a fast determination of transition paths, which drives transition between two known conformations keeping all moment geometries that do not violate the chemical structure and that in general can be explained from the pattern of easiest deformations of the reference structures. In the lack of experimental dynamic data, there is, however, no guarantee on

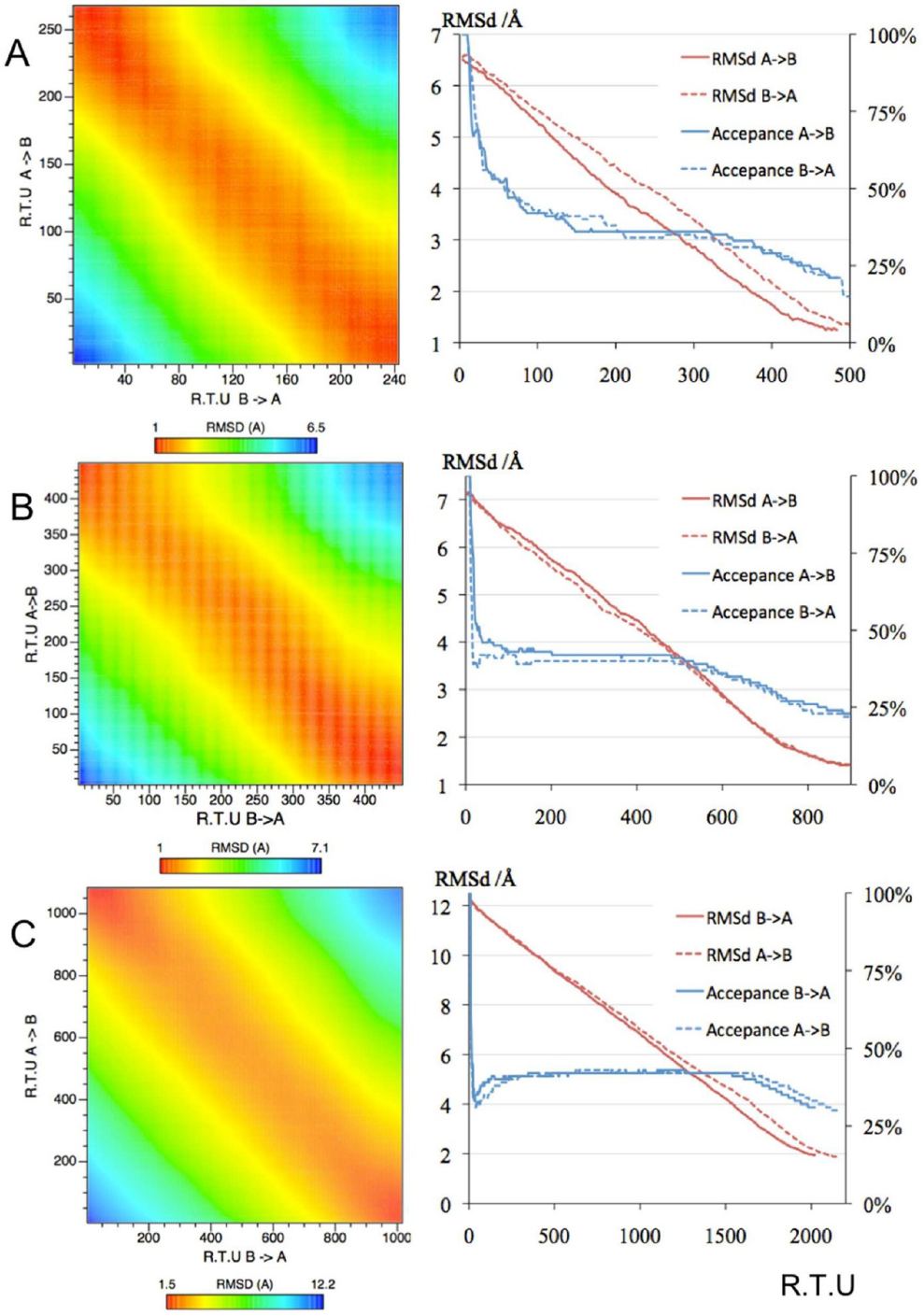


Figure 5. Examples (1l5e ←→ 1l5b, top; 1ake←→4ake, middle; and 1sx4←→1oel, bottom) of transitions obtained by the MDdMD procedure. On the left panel, bidimensional RMSD plots are presented to show the reversibility of the transitions (nearly diagonal distributions). On the right-hand side, Maxwell−Demon acceptance and RMSD profiles along transitions are presented to confirm the reversibility (noted as near-superposition of the curves).

the goodness of the proposed path. Fortunately, for a few cases, PDB reports intermediate structures, i.e., conformations of the protein that are expected to be in the path of transition between two more distant protein conformers (see Figure 6). It must be stressed that there is no direct evidence that the experimental intermediate is a real intermediate for the transition, but it is a reasonable assumption that a good transition path should approach at some point the putative intermediate structure. Analysis of the corresponding transition (Figure 8, Figure S2) illustrates how our transition trajectory typically passes close to the suggested experimental intermediate sampling in all of the case structures, with clear folded-like properties (see Figure 6)

dx.doi.org/10.1021/ct300494q | J. Chem. Theory Comput. 2012, 8, 4707−4718

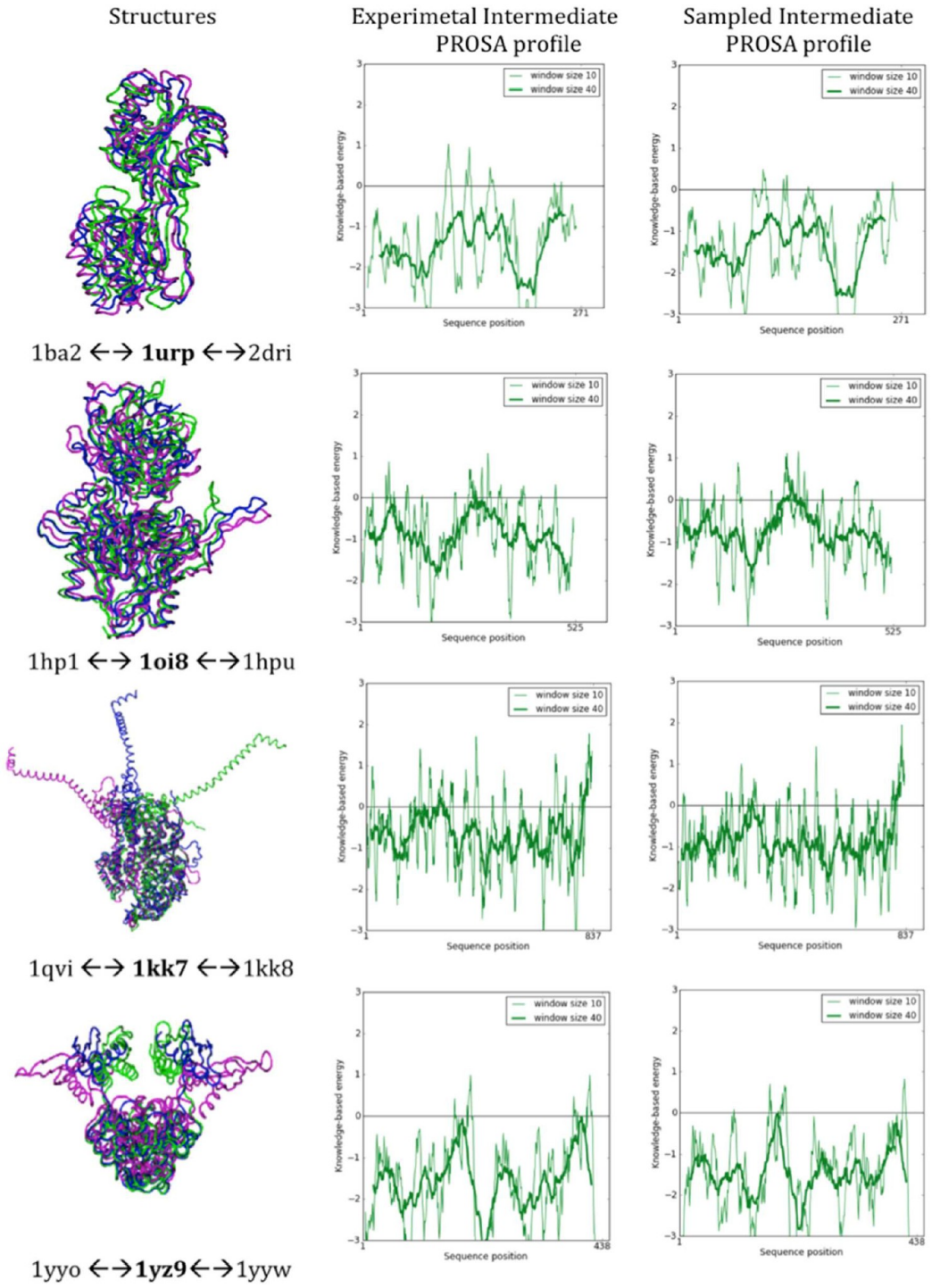


Figure 6. Study of the similarity between experimental and MDdMD transition intermediates. Left: detail of the transitions (with the intermediate displayed in blue) considered here. Right: PROSA profiles for the experimental and MDdMD sampled intermediate (note that no bias was introduced in the simulations to approach MDdMD sampling to experimental intermediate).

and non-negligible RMSDs between the sampled structure and the assumed "intermediate" crystal structure, which are in some cases clearly linked to mobile movements.

Transition Perturbation. The procedure outlined here is very flexible, allowing the introduction of any external effect into the calculation. This allows us to determine, for example, how a given transition might be facilitated or disturbed by the presence of external fields or molecules. Figure 9 illustrates the power of the method to characterize the disturbing effect of the ligand on the close→open transition of adenylate kinase (1ake→4ake), which

dx.doi.org/10.1021/ct300494q | J. Chem. Theory Comput. 2012, 8, 4707−4718

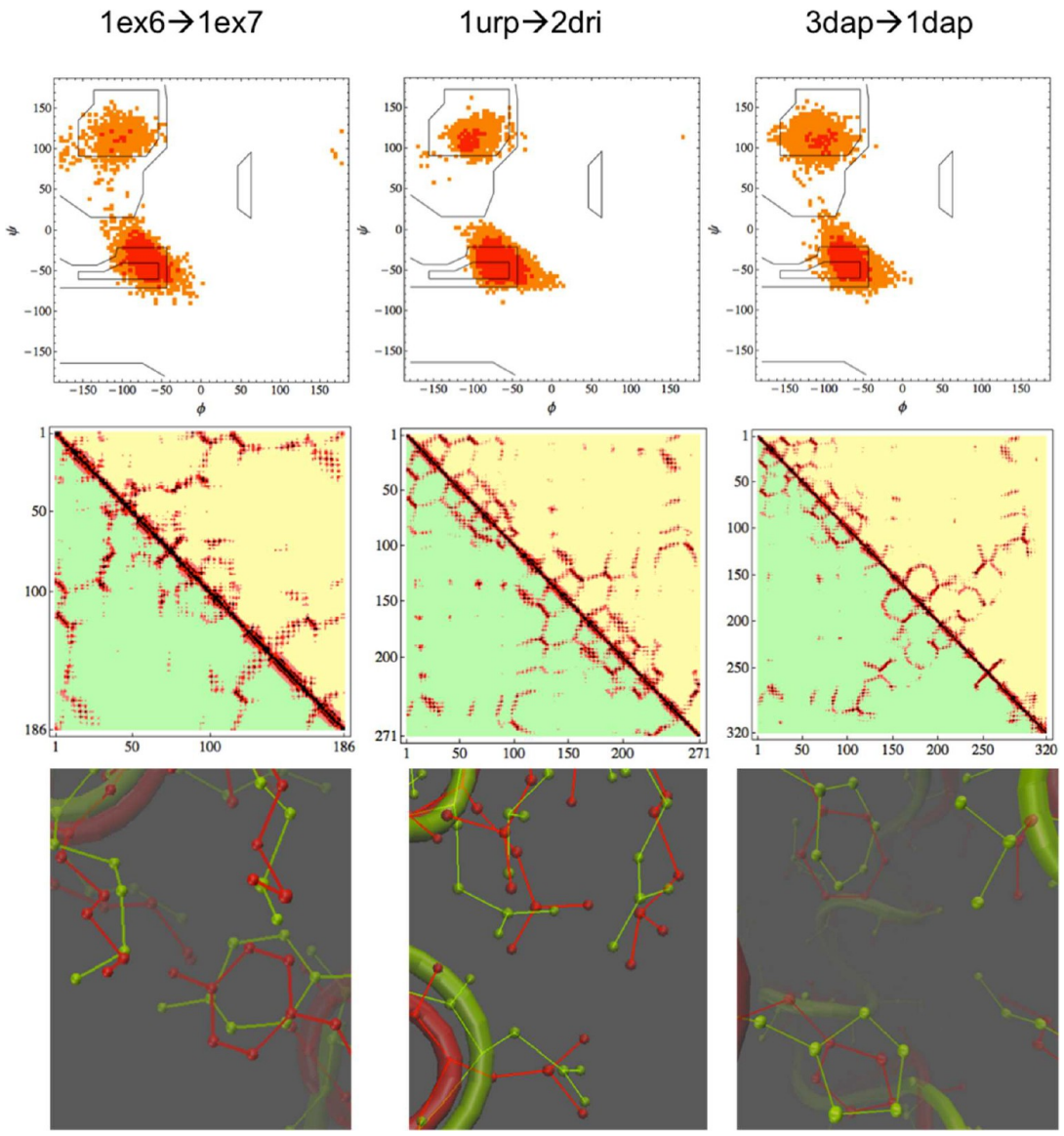


Figure 7. Detail of the local quality of sampled structures for three random cases. Top: Ramachandran maps for all sampled structures (red show higher populated states). Middle: C_β based contact maps for the sampled (above diagonal) and experimental (below diagonal) end state. Bottom: Some structural details of the side chains in the experimental vs final sampled structures.

is characterized by a dramatic decrease in the velocity of the algorithm to advance toward the target structure and the much larger value of RMSD to the target value obtained in the simulations.

MDdMD Application. MDdMD can be run through a Web-based interface (http://mmb.irbbarcelona.org/MDdMD). The user can upload both input and target structures or fetch them from the PDB. Since only Cα from the target structure is used, there is no need for both proteins to be coincident, so mutated structures or close homologues could also be simulated. A reduced set of parameters could also be defined: simulation temperature, acceptance ratio, and the desired final RMSD. Alternatively, most relevant trajectories for the already available simulations (see Table S2) could be examined. After an initial check for the consistency of parameters and structures, simulations are launched. MDdMD simulations are executed on our Web-applications cluster, under an SGE batch queue engine. The simulation can be followed interactively on an intermediate Web page that includes information about the current stage of the transition: an RMSD/acceptance rate plot showing the current acceptance rate and the RMSD between the simulated and the target structure; the superposition between initial, current, and target structures can be visualized through a JMol applet (http://jmol.sourceforge.net). From the intermediate screen, the simulation can be stopped by the user at any time, obtaining the accumulated trajectory as a final result. This helps to avoid nonproductive simulations, or finish simulations that have already reached their target. In any case, the user is informed through an email message of the completion of the simulation, and a link to the final results is provided.

The resulting conformational transition trajectory can be visualized in JMol or downloaded, either in PDB or Amber's

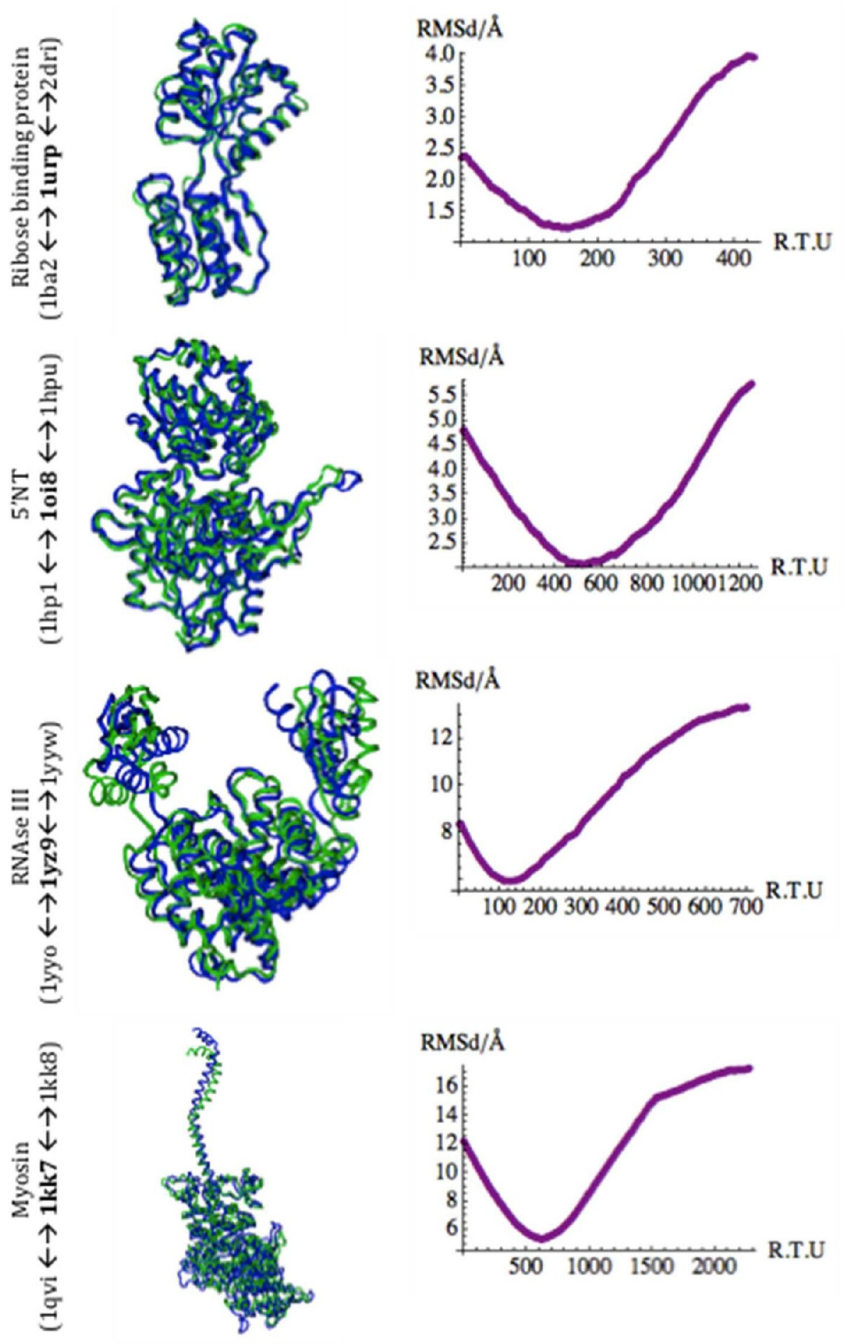


Figure 8. Evolution of the RMSD (to the experimental intermediate) along the simulation (green) together with the experimental intermediate (blue).

ASCII CRD formats, that make them compatible with most MD analysis software.

■ CONCLUSIONS

We present an extension of the discrete molecular dynamics algorithm to trace conformational transitions. The method can work at any level of resolution (including the all-heavy-atoms one explored here) and drives physically meaningful transitions toward the target structure using a Maxwell−Demon procedure enriched by introducing information on the essential deformation pattern of proteins. The method is extremely flexible, allowing the introduction of any desired experimental constraints or the simulation of perturbing elements in the transition. Testing of the method in an extended set of transitions (nearly 100) reveals a success rate around 94%, including some very difficult transitions, involving large movements that do not align with the essential deformability pattern of proteins. The intrinsic speed of dMD makes the technique very efficient computationally and competitive with less physical approaches.

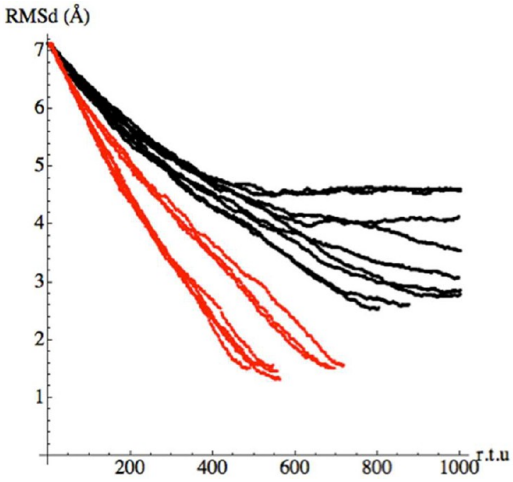


Figure 9. RMSD (to target structure) evolution along simulation time in the 1ake→4ake transition in the absence (red lines) and presence (black lines) of a ligand in the binding site of bis(adenosine)5′-pentaphosphate. The disturbing effect of the ligand is clearly visible in the failure of black lines to reach the target structure. Multiple lines correspond to individual trajectories performed to verify the robustness of the results.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Detailed data of studied conformational transitions. The material is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: modesto.orozco@irbbarcelona.org.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Science and Innovation (BIO2009-10964 and Consolider E-Science), European Research Council (ERC Advanced Grant), Scalalife European Project, and the Fundación Marcelino Botín. P.S is an IRB-la Caixa predoctoral fellow. M.O. is an ICREA Academia investigator.

■ REFERENCES

(1) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hübner, C. G.; Kern, D. *Nature* **2007**, *450*, 838−844.
(2) Velazquez-Muriel, J. A.; Rueda, M.; Cuesta, I.; Pascual-Montano, A.; Orozco, M.; Carazo, J.-M. *BMC Struct. Biol.* **2009**, *9*, 6.
(3) Bakan, A.; Bahar, I. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 14349−14354.
(4) Yang, L.; Song, G.; Jernigan, R. L. *Biophys. J.* **2007**, *93*, 920−929.
(5) Bahar, I.; Chennubhotla, C.; Tobi, D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 633−640.
(6) Tobi, D.; Bahar, I. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 18908−18913.
(7) Eyal, E.; Dutta, A.; Bahar, I. *WIREs Comput. Mol. Sci.* **2011**, *1*, 426−439.
(8) Dobbins, S. E.; Lesk, V. I.; Sternberg, M. J. E. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 10390−10395.
(9) Gerstein, M.; Krebs, W. *Nucleic Acids Res.* **1998**, *26*, 4280−4290.
(10) Falke, J. J. *Science* **2002**, *295*, 1480−1481.
(11) Leo-Macias, A.; Lopez-Romero, P.; Lupyan, D.; Zerbino, D.; Ortíz, A. R. *Biophys. J.* **2005**, *88*, 1291−1299.
(12) Stein, A.; Rueda, M.; Panjkovich, A.; Orozco, M.; Aloy, P. *Structure* **2011**, *19*, 881−889.
(13) Orellana, L.; Rueda, M.; Ferrer-Costa, C.; Lopez-Blanco, J. R.; Chacón, P.; Orozco, M. *J. Chem. Theory Comput.* **2010**, *6*, 2910−2923.
(14) Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128−132.
(15) Ban, D.; Funk, M.; Gulich, R.; Egger, D.; Sabo, T. M.; Walter, K. F. A.; Fenwick, R. B.; Giller, K.; Pichierri, F.; de Groot, B. L.; Lange, O. F.; Grubmüller, H.; Salvatella, X.; Wolf, M.; Loidl, A.; Kree, R.; Becker, S.; Lakomek, N.-A.; Lee, D.; Lunkenheimer, P.; Griesinger, C. *Angew. Chem., Int. Ed.* **2011**, *50*, 11437−11440.
(16) Fenwick, R. B.; Esteban-Martin, S.; Richter, B.; Lee, D.; Walter, K. F. A.; Milovanovic, D.; Becker, S.; Lakomek, N. A.; Griesinger, C.; Salvatella, X. *J. Am. Chem. Soc.* **2011**, *133*, 10336−10339.
(17) Kubitzki, M. B.; de Groot, B. L. *Structure* **2008**, *16*, 1175−1182.
(18) Maragakis, P.; Karplus, M. *J. Mol. Biol.* **2005**, *352*, 807−822.
(19) Shimamura, T.; Weyand, S.; Beckstein, O.; Rutherford, N. G.; Hadden, J. M.; Sharples, D.; Sansom, M. S. P.; Iwata, S.; Henderson, P. J. F.; Cameron, A. D. *Science* **2010**, *328*, 470−473.
(20) Paci, E.; Lindorff-Larsen, K.; Dobson, C. M.; Karplus, M.; Vendruscolo, M. *J. Mol. Biol.* **2005**, *352*, 495−500.
(21) Orozco, M.; Orellana, L.; Hospital, A.; Naganathan, A.; Emperador, A.; Carrillo, O.; Gelpí, J. *Advances in Protein Chemistry and Structural Biology*; Christov, C., Ed.; Burlington Academic Press: Burlington, MA, 2011; Vol. *85*, pp 183−215.
(22) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Pérez, A.; Camps, J.; Hospital, A.; Gelpí, J. L.; Orozco, M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 796−801.
(23) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6679−6685.
(24) Amadei, A.; Linssen, A.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412−425.
(25) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.
(26) Juraszek, J.; Bolhuis, P. G. *Biophys. J.* **2008**, *95*, 4246−4257.
(27) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. *Adv. Chem. Phys.* **2002**, 1−78.
(28) Juraszek, J.; Vreede, J.; Bolhuis, P. G. *Chem. Phys.* **2012**, *396*, 30−44.
(29) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.
(30) Bolhuis, P. G. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 12129−12134.
(31) Wales, D. J. *Int. Rev. Phys. Chem.* **2006**, *25*, 237−282.
(32) Khalili, M.; Wales, D. J. *J. Phys. Chem. B* **2008**, *112*, 2456−2465.
(33) Wales, D. J.; Bogdan, T. V. *J. Phys. Chem. B* **2006**, *110*, 20765−20776.
(34) Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2003**, *119*, 9947−9955.
(35) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585−590.
(36) Brooks, C.; Karplus, M.; Pettitt, M. *Adv. Chem. Phys.* **1988**, *71*, 35−58.
(37) Leone, V.; Marinelli, F.; Carloni, P.; Parrinello, M. *Curr. Opin. Struct. Biol.* **2010**, *20*, 148−154.
(38) Beckstein, O.; Denning, E. J.; Perilla, J. R.; Woolf, T. B. *J. Mol. Biol.* **2009**, *394*, 160−176.
(39) Perilla, J. R.; Beckstein, O.; Denning, E. J.; Woolf, T. B. *J. Comput. Chem.* **2010**, *32*, 196−209.
(40) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(41) Barducci, A.; Bonomi, M.; Parrinello, M. *WIREs Comput. Mol. Sci.* **2011**, *1*, 826−843.
(42) Schlitter, J. *J. Mol. Graph.* **1995**, *12*, 84−89.
(43) Krüger, P.; Verheyden, S.; Declerck, P. J.; Engelborghs, Y. *Protein Sci.* **2001**, *10*, 798−808.
(44) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690−2693.
(45) Liphardt, J.; Dumont, S.; Smith, S. B.; Tinoco, I.; Bustamante, C. *Science* **2002**, *296*, 1832−1835.

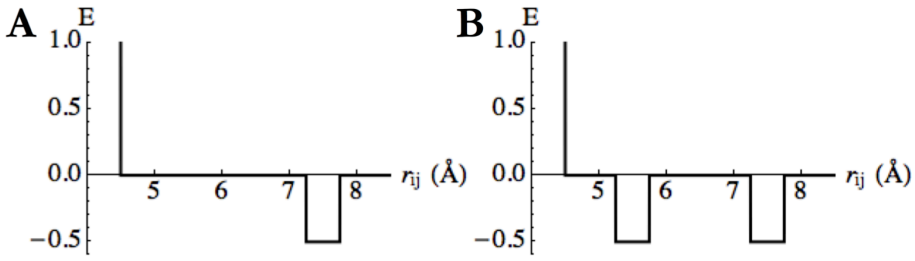
(46) Bahar, I.; Rader, A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586−592.
(47) Derreumaux, P.; Mousseau, N. *J. Chem. Phys.* **2007**, *126*, 025101.
(48) Tirion, M. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.
(49) Devane, R.; Shinoda, W.; Moore, P. B.; Klein, M. L. *J. Chem. Theory Comput.* **2009**, *5*, 2115−2124.
(50) Kim, M. K.; Chirikjian, G. S.; Jernigan, R. L. *J. Mol. Graphics Modell.* **2002**, *21*, 151−160.
(51) Kim, M. K.; Jernigan, R. L.; Chirikjian, G. S. *Biophys. J.* **2002**, *83*, 1620−1630.
(52) Mendez, R.; Bastolla, U. *Phys. Rev. Lett.* **2010**, *104*, 228103.
(53) Lopez-Blanco, J. R.; Garzón, J. I.; Chacón, P. *Bioinformatics* **2011**, *27*, 2843−2850.
(54) Rueda, M.; Chacón, P.; Orozco, M. *Structure* **2007**, *15*, 565−575.
(55) Ding, F.; Dokholyan, N. V. *Trends Biotechnol.* **2005**, *23*, 450−455.
(56) Ding, F.; Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *Biophys. J.* **2002**, *83*, 3525−3532.
(57) Shirvanyants, D.; Ding, F.; Tsao, D.; Ramachandran, S.; Dokholyan, N. V. *J. Phys. Chem. B* **2012**, *116*, 8375−8382.
(58) Emperador, A.; Meyer, T.; Orozco, M. *J. Chem. Theory Comput.* **2008**, *4*, 2001−2010.
(59) Emperador, A.; Meyer, T.; Orozco, M. *Proteins: Struct., Funct., Bionf.* **2010**, *78*, 83−94.
(60) Emperador, A.; Carrillo, O.; Rueda, M.; Orozco, M. *Biophys. J.* **2008**, *95*, 2127−2138.
(61) Ding, F.; Buldyrev, S.; Dokholyan, N. *Biophys. J.* **2005**, *88*, 147−155.
(62) Zhou, Y.; Karplus, M. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 14429−14432.
(63) Ding, F.; Sharma, S.; Chalasani, P.; Demidov, V. V.; Broude, N. E.; Dokholyan, N. V. *RNA* **2008**, *14*, 1164−1173.
(64) Krebs, W. G.; Gerstein, M. B. *Nucleic Acids Res.* **2000**, *28*, 1665−1675.
(65) Ye, Y.; Godzik, A. *Nucleic Acids Res.* **2004**, *32*, W582−W585.
(66) Flores, S.; Echols, N.; Milburn, D.; Hespenheide, B.; Keating, K.; Lu, J.; Wells, S.; Yu, E. Z.; Thorpe, M.; Gerstein, M. *Nucleic Acids Res.* **2006**, *34*, D296−D301.
(67) Lindahl, E.; Azuara, C.; Koehl, P.; Delarue, M. *Nucleic Acids Res.* **2006**, *34*, W52−W56.
(68) Franklin, J.; Koehl, P.; Doniach, S.; Delarue, M. *Nucleic Acids Res.* **2007**, *35*, W477−W482.
(69) Weiss, D. R.; Levitt, M. *J. Mol. Biol.* **2009**, *385*, 665−674.
(70) Yang, Z.; Májek, P.; Bahar, I. *PLoS Comput. Biol.* **2009**, *5*, e1000360.
(71) Lezon, T. R.; Sali, A.; Bahar, I. *PLoS Comput. Biol.* **2009**, *5*, e1000496.
(72) Bahar, I.; Lezon, T. R.; Yang, L.-W.; Eyal, E. *Annu. Rev. Biophys.* **2010**, *39*, 23−42.
(73) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. *Chem. Rev.* **2010**, *110*, 1463−1497.
(74) Bryngelson, J.; Onuchic, J.; Socci, N.; Wolynes, P. *Proteins: Struct., Funct., Bionf.* **1995**, *21*, 167−195.
(75) Proctor, E. A.; Ding, F.; Dokholyan, N. *WIREs Comput. Mol. Sci.* **2011**, *1*, 80−92.
(76) Ding, F.; Tsao, D.; Nie, H.; Dokholyan, N. V. *Structure* **2008**, *16*, 1010−1018.
(77) Urbanc, B.; Borreguero, J.; Cruz, L.; Stanley, H. *Methods Enzymol.* **2006**, *412*, 314−338.
(78) Smith, S.; Hall, C.; Freeman, B. *J. Comput. Phys.* **1997**, *134*, 16−30.
(79) Taketomi, H.; Ueda, Y.; Gō, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445−459.
(80) Camps, J.; Carrillo, O.; Emperador, A.; Orellana, L.; Hospital, A.; Rueda, M.; Cicin-Sain, D.; D'Abramo, M.; Gelpí, J. L.; Orozco, M. *Bioinformatics* **2009**, *25*, 1709−1710.
(81) Nguyen, H. D.; Hall, C. K. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 16180−16185.
(82) Gherghe, C. M.; Leonard, C. W.; Ding, F.; Dokholyan, N. V.; Weeks, K. M. *J. Am. Chem. Soc.* **2009**, *131*, 2541−2546.
(83) Hajdin, C. E.; Ding, F.; Dokholyan, N. V.; Weeks, K. M. *RNA* **2010**, *16*, 1340−1349.
(84) Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *Folding Des.* **1998**, *3*, 577−587.
(85) Ding, F.; Lavender, C. A.; Weeks, K. M.; Dokholyan, N. V. *Nat. Methods* **2012**, 603−608.
(86) Peng, S.; Ding, F.; Urbanc, B.; Buldyrev, S. V.; Cruz, L.; Stanley, H. E.; Dokholyan, N. V. *Phys. Rev. E* **2004**, *69*, 041908.
(87) Urbanc, B.; Betnel, M.; Cruz, L.; Bitan, G.; Teplow, D. *J. Am. Chem. Soc.* **2010**, *132*, 4266−4280.
(88) Urbanc, B.; Cruz, L.; Yun, S.; Buldyrev, S. V.; Bitan, G.; Teplow, D. B.; Stanley, H. E. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17345−17350.
(89) Ding, F.; Borreguero, J.; Buldyrev, S. V.; Stanley, H.; Dokholyan, N. *Proteins: Struct., Funct., Bionf.* **2003**, *53*, 220−228.
(90) Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Bionf.* **1999**, *35*, 133−152.
(91) Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Pérez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpí, J. L.; Orozco, M. *Structure* **2010**, *18*, 1399−1409.
(92) Rueda, M.; Cubero, E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2004**, *87*, 800−811.
(93) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(94) Wiederstein, M.; Sippl, M. J. *Nucleic Acids Res.* **2007**, *35*, W407−W410.

Chapter 4: Speeding up the transition path sampling

In the previous chapter we introduced an algorithm to atomistically connect two known conformers of the same protein. Here, we present an alternative method that is at the same time more efficient and more accurate.

We needed a way to test a large number of plausible conformational transitions paths, first with two known end points, but aiming to predict transitions paths. Despite the efficiency of MDdDM, it was still too slow to systematically test, for instance, 1000 conformational transitions per protein. The Cα resolution was adequate to obtain such computational efficiency, and we adopted a SBM inspired by its success to describe protein conformational flexibility. Firstly, in our group, Orellana et al showed that disease mutations correlate with protein dynamics disrupting Cα interactions in a ENM (*129*). Other success stories from SBM models come are the description of the mechanism of action of Adenylate Kinase (*112*) or the concerted motion of two leading heads of Kinesin (*243*). We benefited from dMD versatile square wells to build a double-minima potential energy landscape using two known structures. The protocol can be easily extendable to add more minima or incorporate restrains. Figure 7 shows an example or energy potential used in this model.

Figure 7: Double-well square potentials



Modeling conformational changes with simple SBM potentials. For pair of particles that do not change their relative distance (r_{ij}) in the conformational transition, a single energy minimum is placed at the reference distance of both known structures (A). When the distance between particles changes along with the conformational transition two minima are used (B). Minima position correspond to each distance in the end conformers, typically two PDB structures.

We observed that with this simple model we are able to sample spontaneously known intermediates lying along the transition path at accuracy below 2 Å RMSD for medium sizes proteins.

Title: **Exploration of Conformational Transition Pathways from Coarse-grained Simulations**

Authors: Pedro Sfriso, Adam Hospital, Agustí Emperador, and Modesto Orozco

Stage: Published

Journal: Bioinformatics

Type: Research Article

Supplementary Material:
http://bioinformatics.oxfordjournals.org/content/early/2013/06/05/bioinformatics.btt324/suppl/DC1

Author Contribution: PS was the main responsible all the work, developed the method and ran the simulations. PS contributed to the writing the paper.

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 16 2013, pages 1980–1986
doi:10.1093/bioinformatics/btt324

Structural bioinformatics

Advance Access publication July 5, 2013

Exploration of conformational transition pathways from coarse-grained simulations

Pedro Sfriso[1,2], Adam Hospital[1,2,3], Agustí Emperador[1,2] and Modesto Orozco[1,2,3,4,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), [2]Joint IRB-BSC Program in Computational Biology, Baldiri Reixac 10, Barcelona 08028, Spain, [3]National Institute of Bioinformatics and [4]Department of Biochemistry and Molecular Biology, University of Barcelona, Av. Diagonal 647, Barcelona 08028, Spain

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: A new algorithm to trace conformational transitions in proteins is presented. The method uses discrete molecular dynamics as engine to sample protein conformational space. A multiple minima Go-like potential energy function is used in combination with several enhancing sampling strategies, such as metadynamics, Maxwell Demon molecular dynamics and essential dynamics. The method, which shows an unprecedented computational efficiency, is able to trace a wide range of known experimental transitions. Contrary to simpler methods our strategy does not introduce distortions in the chemical structure of the protein and is able to reproduce well complex non-linear conformational transitions. The method, called GOdMD, can easily introduce additional restraints to the transition (presence of ligand, known intermediate, known maintained contacts, ...) and is freely distributed to the community through the Spanish National Bioinformatics Institute (http://mmb.irbbarcelona.org/GOdMD).

Availability: Freely available on the web at http://mmb.irbbarcelona.org/GOdMD.

Contact: modesto.orozco@irbbarcelona.org or modesto@mmb.pcb.ub.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 26, 2013; revised on April 29, 2013; accepted on May 30, 2013

1 INTRODUCTION

Many biological functions of proteins such as mechanic work, signal transduction or enzymatic activity are modulated by a key property of them: flexibility (Henzler-Wildman *et al*., 2007; Micheletti, 2013; Velazquez-Muriel *et al*., 2009). Flexibility is a property that has been refined and maintained by evolution (Falke, 2002; Micheletti, 2013) and that, in turn has been also exploited by evolution to generate new proteins in a conservative mechanism, which guarantees the maintenance of the structural scaffold as well as the relevant deformation pattern (Leo-Macias *et al*., 2005; Stein *et al*., 2011). Structural databases show (Gerstein and Krebs, 1998) increasing number of proteins having alternative structures depending on external factors (such as crystallization conditions, posttransductional chemical modifications, presence of ligands, changes in solvent environment, etc). This probes the existence of dramatic conformational transitions in proteins, but giving no information on how such transitions happen.

Recent refinements of experimental techniques have provided direct evidence on the mechanisms of conformational transitions for some model proteins (Ban *et al*., 2011; Eisenmesser *et al*., 2002; Fenwick *et al*., 2011; Kern and Zuiderweg, 2003; Lindorff-Larsen *et al*., 2005). However, we are still far from the point that all conformational transitions could be described by means of experimental methods. This situation forces the use of simulation techniques, which has been largely refined in the past years (Best *et al*., 2005; Bolhuis *et al*., 2002; Karplus and Kuriyan, 2005; Kubitzki and de Groot, 2008; Maragakis and Karplus, 2005; Miyashita, 2003; Okazaki *et al*., 2006), providing information of increasing quality for a large number of conformational transitions in proteins (Sfriso *et al*., 2012; Stein *et al*., 2011).

Molecular dynamics (MD), using atomistic force-fields and explicit representation of solvent (McCammon *et al*., 1977), is probably the most accurate theoretical technique for reproducing protein flexibility. Recent computational approaches have made possible the simulation of thousands of proteins (Meyer *et al*., 2010) and the derivation of up to millisecond trajectories for proteins (Dror *et al*., 2012; Lindorff-Larsen *et al*., 2011). Unfortunately, most conformational transitions are still far from the capabilities of plain atomistic MD, forcing the use of biasing schemes (Beckstein *et al*., 2009; Das *et al*., 2006; Laio and Parrinello, 2002; Leone *et al*., 2010; Liphardt *et al*., 2002; Schlitter, 1994), designed to maximize the sampling along a given variable that is believed to capture conformational transition motions. Biased-MD protocols are extremely powerful, but they are expensive computationally, require expertise from the user and can lead to incorrect results when the transition coordinate is not well defined.

Coarse-grained (CG) models (Bahar and Rader, 2005; Dobbins *et al*., 2008; Lopez-Blanco *et al*., 2011; Mendez and Bastolla, 2010; Orozco *et al*., 2011; Tozzini, 2005; Whitford *et al*., 2007) are inexpensive, but still accurate alternatives to atomistic MD simulations, which have gained significant popularity in recent years. Unfortunately, the use of CG models require the assumption of a certain loss of detail in the simulation, for example, explicit solvent is ignored, which prevent the study of specific water–protein interactions and side chains are either ignored or dramatically simplified (Mendez and Bastolla, 2010; Marrink *et al*., 2007), which raises problems

*To whom correspondence should be addressed.

to reproduce ligand-target interactions. CG models have been calibrated against experimental data, such as B-factors, structural variability in databases or atomistic MD, and despite their simplicity, provide often results of surprising quality with limited computer resources and reduced expertise from the user.

CG-methods have been largely used to trace conformational transitions in proteins. The first implementations followed mainly interpolation schemes between original and final conformations (Delarue and Sanejouand, 2002; Flores *et al*., 2006; Franklin *et al*., 2007; Kim *et al*., 2002; Krebs and Gerstein, 2000; Lindahl *et al*., 2006; Weiss and Levitt, 2009). Interpolation protocols are fast and guarantee the completion of the transition, but often at the expense of unrealistic intermediate conformations, which are not good starting points for refinement through more accurate atomistic simulations. Alternative CG-transition methods have been developed under the assumption that biologically relevant transitions should follow the easiest deformation movements of the proteins (Bahar *et al*., 2010; Lezon *et al*., 2009; Mendez and Bastolla, 2010; Yang *et al*., 2009) defined as the softest deformation modes obtained by diagonalization of the Hessian matrix derived from an elastic network model (ENM) Hamiltonian:

$$E = \sum_{i,j} \delta_{ij} K_{ij} \left(R_{ij} - R_{ij}^{0} \right)^{2} \quad (1)$$

where i and j are residues, δ_{ij} is a delta function equal to 1 when i and j are at less than a given distance, and 0 otherwise; K is a spring constant (linear or distance dependent), R_{ij} stands for inter-residue distance and the superscript 0 refers to the value of K_{ij} in the reference structure.

Morphing methods based on the use of essential deformation modes provide more reasonable approaches than linear interpolation schemes, but are far from being optimal since (i) the covalent structure of the protein can be damaged when large displacements along a limited number of Cartesian eigenvectors are made, (ii) a non-negligible number of transitions do not align with the intrinsic deformation pattern of proteins (Stein *et al*., 2011), and even those cases where the alignment between the essential deformation space and the transition vector is good, displacement along the essential space of one of the proteins never allows a full transition between the two conformational states. Recent advances in the field, such as the use of multiple reference structures to define a transition-dependent essential deformation space (Sfriso *et al*., 2012; Yang *et al*., 2009) or the derivation of the essential deformation modes in the dihedral space (Lopez-Blanco *et al*., 2011; Mendez and Bastolla, 2010), have alleviated but not solved these problems.

In this article, we present a new approach to obtain ultra-fast, but accurate, conformational transition pathways in proteins out of physics-based molecular mechanics simulations. By combining an efficient sampling technique [discrete molecular dynamics (dMD; Orozco *et al*., 2011; Proctor *et al*., 2011)] with a novel multiple-well Go-like scheme. Sampling is enhanced with biasing techniques such as metadynamics (Laio and Parrinello, 2002) and Maxwell Demon MD (Rueda *et al*., 2004; Sfriso *et al*., 2012). The method was tested in a large battery of transitions (near 50 pairs of structures) obtaining, with reduced computational cost, reasonable pathways in all cases, including in cases of extreme difficulty, where transition requires, for example, partial unfolding of the protein. Owing to the physical nature of the method, sampled intermediate structures maintain covalent structure, and no steric clashes are allowed. In cases where experimental intermediates are characterized, the method finds transition paths passing close to them. Surprisingly, the method outperforms our previous more detailed algorithm, Maxwell Demon discrete molecular dynamics procedure (MDdMD; Sfriso *et al*., 2012), and is also competitive with respect to alternative methods in the literature (Supplementary Figs S1–S3). Our novel approach is flexible, allowing the introduction of any perturbation. Furthermore, the user can easily bias trajectories to guarantee that known experimental intermediates are sampled.

2 METHODS

In this approach we have applied the dMD method [see (Emperador *et al*., 2008b; Proctor *et al*., 2011) for details]. In dMD, particles move in the ballistic regime, with constant velocity until a collision occurs. Collisions occur at the particle–particle distances corresponding to a discontinuity in the dMD potential. The velocities of the particles after the collision are obtained using the rules of conservation of momentum and energy. This allows avoiding the integration of Newton's equations of motion, speeding up the simulations as compared with usual MD. Discrete MD has successfully been applied to protein folding (Ding *et al*., 2005; Zhou and Karplus, 1997), macromolecular dynamics (Emperador *et al*., 2008a, 2010), RNA structural predictions (Ding *et al*., 2008, 2012; Gherghe *et al*., 2009) and protein aggregation (Ding and Dokholyan, 2008; Urbanc *et al*., 2004, 2010). Recently dMD has been also used to robust energy minimization of protein–protein complexes (Emperador *et al*., 2013) and protein–ligand interactions (Proctor *et al*., 2012).

2.1 Force-field representation

We adopted a CG representation of the protein, where only Cαs were explicitly represented. We used a single square well with infinite walls to define the bonded interactions between consecutive beads. The non-bonded interactions were described by a multiple well Go-like model (Taketomi *et al*., 1975; Ueda *et al*., 1978) depending on the experimental inter-particle distance. Non-bonded terms of the Hamiltonian were classified in two categories based on whether or not the inter-particle distance is different between the initial (A) and final (B) conformations. If they are, we define a double well potential (Fig. 1) centered in the respective experimental values: $\vec{r}_A$ and $\vec{r}_B$. When the distances are similar in both conformations (i.e. $\vec{r}_A \sim \vec{r}_B$), we used a single, but wider, well (50% larger) centered at ½ ($\vec{r}_A + \vec{r}_B$). To avoid over-restraining the system with physically irrelevant interactions, the potential energy was defined only for those particles within a cutoff ($\vec{r}_A$, $\vec{r}_B >$ radii of gyration/2). It is worth noting that the height of the wells for all non-bonded interactions is finite (0.5 kcal/mol), allowing the particles to escape from the wells if required.

This Go-like scheme is convenient to complete transition paths [>85% of the root mean square deviation (RMSd) difference between end structures) because target minimum acts as attractor. This situation enables us to use softer biasing schemes and to better recover native contacts (see Supplementary Materials for details].

2.2 Accelerating the transition

Within the Hamiltonian definition above, the A→B transition occurs through a movement of particles jumping from wells centered at $\vec{r}_A$ to those centered at $\vec{r}_B$. Unfortunately, as a side effect of Go-like potentials,

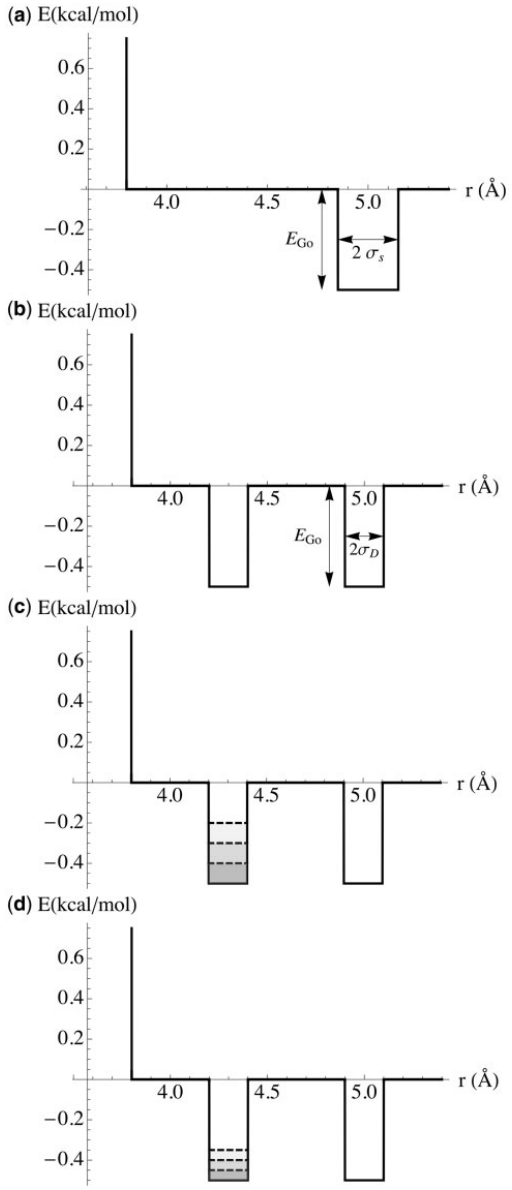


Fig. 1. Interaction potentials used in this study. (**a**) single well corresponding to a particle–particle non-bonded pair that does not change their distance during transition ($E_{GO} = 0.5$ kcal/mol, $\sigma_s = 0.15$ Å). (**b**) Double well representing two states of particle radial distance that corresponds to two known structures ($E_{GO} = 0.5$ kcal/mol, $\sigma_D = 0.10$ Å). (**c**) Example of how discrete metadynamics increases the potential energy of a pair of particles whose movement overlaps with the essential deformation space of the initial state and (**d**) the same situation but when the overlap is much lower

spontaneous hopping from one minimum to another is difficult, making unlikely to find a spontaneous transition. We decided then to bias the trajectory to guarantee enrichment of sampling along the transition path. For this purpose, we implemented here a complex biasing scheme designed to favor the transition, but avoiding the use of restrains that would guide the trajectory along arbitrary (and probably unrealistic paths). The first level of biasing is designed to move the protein far from the starting conformation, while the second level is designed to enrich biasing to approach the protein to the final conformation. For the interested reader, a flowchart of the algorithm is presented in Supplementary Figure S4.

2.2.1 Escaping from the initial minima We have implemented a discrete version of the metadynamics method (Laio and Parrinello, 2002) to guarantee that the system leaves the original minima. Metadynamics penalizes visits to the original well by raising gradually the energy of the well (i.e. by filling it; Fig. 1), which increases the chances of the system to escape from the attraction of the Go-potentials to the starting structure. As a consequence, as trajectory progresses sampling of the original conformation is less and less probable and system departs from the original conformation. Metadynamics is efficient as a method to escape from a minimum, but such a divergence happens in a random way, which can yield to unrealistic deformations when applied in an unsupervised fashion. To solve this problem, we have coupled metadynamics to an ENM (Emperador *et al.*, 2008a; Orellana *et al.*, 2010) in such a way that wells associated to inter-particle distances showing large changes along the first five essential deformation modes were filled faster than those associated to inter-particle distances that are not coupled to the essential deformation modes of the protein (Fig. 1). This strategy guarantees the exploration of alternative ways to escape from the original structure minima, while increasing the possibility to sample preferred pathways as defined by the essential deformation modes (first five are considered by default, but results are largely invariant in the range 3–10 modes (data not shown). Note that the ENM analysis is done only with the initial structure irrespectively of which the target structure is, avoiding the definition of too linear pathways. Note also that the full dMD simulations are done, which means that all the degrees of freedom (and not only the five preferred ones) are sampled.

2.2.2 Moving toward the final conformation The ENM-metadynamics procedure outlined above guarantees that the protein moves apart from the original structure sampling preferentially essential deformation modes. However, there is no guarantee that such movements will approach the protein to the final conformation. To guide the trajectory toward the final structure, without introducing arbitrary energy restraints, we have implemented a Maxwell Demon biasing algorithm (Perilla *et al.*, 2010; Rueda *et al.*, 2004; Sfriso *et al.*, 2012) with a control magnitude (Γ) defined as follows:

$$\Gamma = \sum_{i=1}^{N} \omega(i) \left\| r_{i,B} - r_{i,X} \right\| \tag{2}$$

where N is the total number of residues, B is the target structure, X is the sampled conformation ($X = A$ for the original conformation) and $\omega(i)$ is an optimized weight function dependent on the inter-particle distance and the size of protein (Supplementary Fig. S5; results are robust to ω values in the range 15–25 Å are used).

Following the MDdMD procedure the bias toward the target structure is not introduced by an energy penalty, but using a less interfering informational criteria (Perilla *et al.*, 2010). The scheme is simple and efficient, after a certain number of dMD simulation steps (time t) the value of the progress variable (Γ; equation 2) is compared with that obtained in a previous accepted movement (t-Δt). The simulation fragment t-$\Delta t \rightarrow t$ is then accepted or rejected based on a simple Metropolis test (equation 3):

$$p_t = \begin{cases} 1 & if \quad \Gamma_t < \Gamma_{t-\Delta t} \\ \exp\left[-\left(\frac{\Gamma_t - \Gamma_{t-\Delta t}}{\beta RMSd_{(X;t,\,B)}} \right)^2 \right] & if \quad \Gamma_t > \Gamma_{t-\Delta t} \end{cases} \tag{3}$$

where p is the acceptance probability, $RMSd_{(X;t,B)}$ is RMSd between structure sampled at time $t(X)$ and target structure, β is dynamically

adjusted to guarantee an acceptation rate of 70%, and the time frame (Δt) is typically 100 time steps. These large acceptance rates, combined with a weight function with a maximum at ~15 Å enables the system to explore no so direct pathways that could lead to stressed structures and provides the best local geometries (Supplementary Figs S6 and S7). The presence of the Go-potentials of the target structure in the Hamiltonian avoids the need to increase β when trajectory approaches the final target conformation.

3 RESULTS

To evaluate the power of GOdMD, we apply it to trace transitions between known equilibrium conformations of proteins. After analyzing protein data bank (PDB), we define a set of 94 transitions corresponding to 47 proteins showing two distinct conformations. This large benchmark set was extracted from a previous work in our group (Sfriso *et al*., 2012) and spans a wide range of proteins from 100 to 1000 residues, showing different shapes and secondary structure composition. The database contains no trivial conformational changes and some of the conformational transitions represent dramatic geometrical alterations in the structure of the protein, including refolding in some cases. Many of the transitions are coupled to the binding to small ligands (60%) and/or macromolecules (39%), which increase the difficulties to trace reasonable conformational transition pathways. Further analysis of the data set reveals that a significant fraction of the transitions (44%) correspond to open/close ones, which can a priori generate hysteresis problems. Finally, 22% of the transitions do not align well [overlap (accumulated dot product) between top five normal modes and conformational transition is below 0.25] with the first five essential deformation modes detected from EN-normal mode analysis (NMA), i.e. that they will be very difficult to represent by simply activating movements along intrinsic deformation modes. Looking at different criteria (Sfriso *et al*., 2012) we consider 62 of the 94 transitions as difficult or very difficult to follow, which means that we are validating our method with the most exigent transition dataset available.

GOdMD has been able to find reasonable pathways for all studied transitions, even those requiring large refolding processes. The algorithm can provide multiple transition trajectories, defining a scenario of multiple pathways, which fits better into the transition funnel theory (Dill and Chan, 1997; Portella and Orozco, 2010). Furthermore, no violations of covalent distances, nor steric clashes or chemical meaningless conformations are sampled during transitions, which are always smooth, without the presence of apparent discontinuities or hysteresis effects (Supplementary Fig. S8). PROSA (Wiederstein and Sippl, 2007) calculations performed in random conformations sampled during the different transition yields to native-like profiles (Supplementary Figs S9 and S10), indicating that we are not sampling unrealistic conformations along the transition. The method is extremely efficient, >60% of trajectories are finished in <2 min wall-clock time in a laptop computer (2.4 GHz Intel Core 2 Duo) and even the most difficult transitions are finished in <30 min in the same computer. We are working in a parallel algorithm to explore larger systems with similar efficiency.

3.1 Sampling known intermediate conformations

There are a few cases in PDB of distinct conformations for a protein, where there is in addition to the start and end conformation a third structure, which is intermediate (at least in terms of RMSd) between the other two. Following Weiss and Levitt (2009) we can suggest that in general the third structure can be near a preferred passing point on the transition between the extreme conformations (i.e. we can consider an 'intermediate' in the transition). Thus, for the five cases where this putative intermediate is available, we determined transition paths between the two more diverse conformations. In all cases (Fig. 2) we found smooth and reasonable transitions, which go in the direction of the putative intermediate, even for the most difficult cases, where simpler methods have serious difficulties. In small systems (Ribose Binding Protein, 5′-NT), the transition passes through the putative intermediate. In larger systems, small RMSd deviations to the putative intermediate are much more difficult to obtain, partly because of noise introduced by the presence of

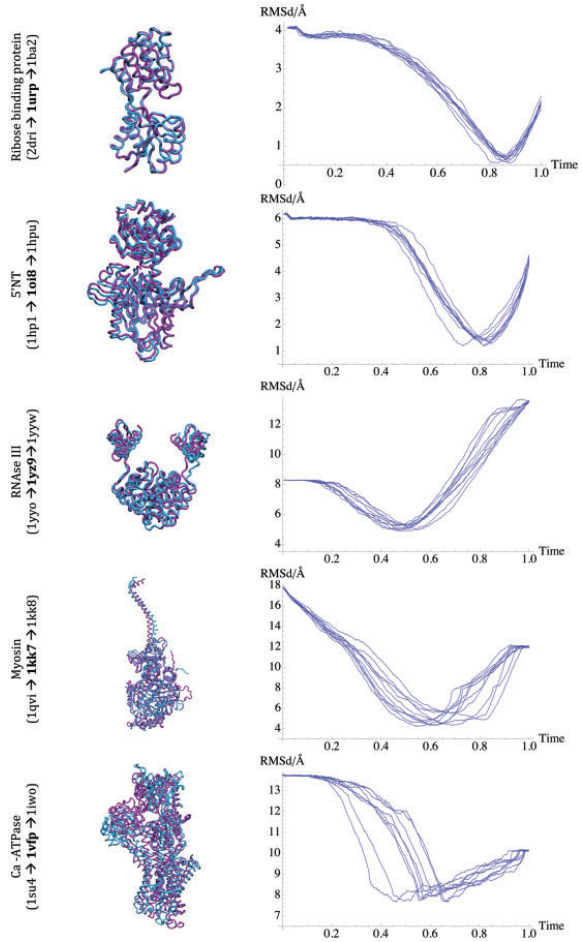


Fig. 2. Structural superposition of experimental intermediate structure and sampled conformations obtained along transitions. In the right-hand side column we display the RMSd profile (taking intermediate as reference) obtained in 10 independent transition pathways

long connecting loops, and partly because the intrinsic difficulty of tracing large conformational transitions. Although some large RMSd values are obtained (due mostly to stochastic loop movements), the potentials used in GOdMD capture better intermediate states than other state-of-the-art morphing methods (Supplementary Fig. S3). However, it is clear that the simulation drives spontaneously the transition toward conformational regions close to the putative intermediate. PROSA profiles and TMScore (Zhang and Skolnick, 2004) of both sampled and putative experimental intermediates are similar (Supplementary Figs S9 and S10 and Supplementary Table S2 for details), supporting the quality of the transition path.

3.2 Non-linearity of the transitions

One of the main caveats of biasing methods based on forcing a regular reduction of RMSd to the target structure (as most morphing methods) is that they force a linear transition path, that might be unrealistic (Fig. 3), and that does not provide information on the bottlenecks in the transition. Our method is able to detect non-linear transition pathways, and provides information on where are the bottlenecks, i.e. those points where the transition seems nearly stopped for long periods, and which correspond to regions of high rejection rate in the Maxwell Demon (Supplementary Fig. S11). Interestingly, the method is also able to disconnect local (measured as % of native contacts) to global transitions (measured by the RMSd to the target), showing how for some transitions local rearrangements happen first, while, on the contrary, for other transitions global movement is before local conformational rearrangements (Figs 3 and 4).

It is worth noting that, the physical nature of the method guarantees that the guiding engine (NMA-bias metadynamics and a Maxwell Demon here) finds conformational pathways that never explore regions of unrealistically large energy. This guarantees that along all the transitions we sample protein-like conformations (Fig. 4), avoiding sampling always linear pathways that can leads to strong local distortions.

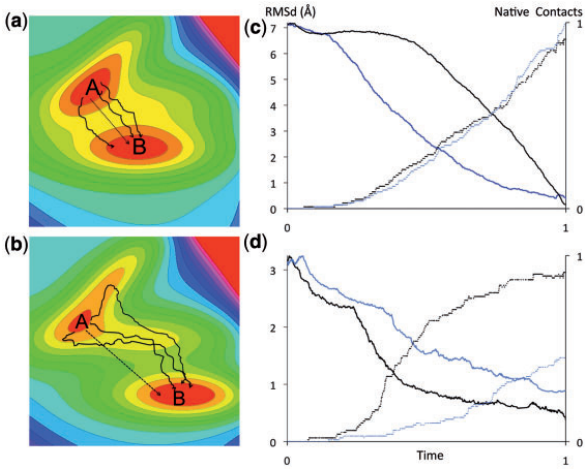


Fig. 3. (**a**) Energy surface where a linear conformational transition is expected to be realistic (x and y correspond to two general conformational variables, and color code means the energy order: red<yellow<green<blue). (**b**) Energy surface where transition does not follow a linear pathway. (**c**) An example of a real transition (4ake→1ake) that behaves linearly, and where pathway obtained by our current procedure (black line) and a linear one (blue; obtained by forcing low acceptance rates and ignoring ENM analysis). (**d**) As in panel (**c**), but for a non-linear transition 1c9kB → 1cbuB, where normal (black) and forced linear transitions (blue) are different, the latter being incorrect, because little target-native contacts are recovered (despite the low RMSd to target)

3.3 Introducing experimental information

One of the main advantages of our method is its flexibility, which makes very easy to introduce the effect of external perturbations in the transition, or to bias the trajectory by experimental information.

To show the capabilities of our algorithm, we analysed the effect of ligand binding in a transition, and also illustrate how the existence of experimentally validated intermediate can be used to bias the transition towards preferred pathways. For the first purpose we studied three additional systems for which ligand binding is known to displace conformational equilibrium: D-Allose binding protein (1gud/1rpj), L-Leucine Binding Protein (1usg/1usi) and Osmo-protection protein (1sw5/1sw2). After running 10 simulations (five in the absence and five in the presence of the ligand), we observed that those trajectories containing ligand were less efficient to reach the final state (Supplementary Figs S12 and S13 for details). These are just 'prove of concept' calculations, but provide encouraging evidence that our simple and fast algorithm can capture qualitatively the effect of the ligand in altering conformational pathways. To explore the possibility to bias GOdMD transitions by using experimental information on intermediate we have repeated the study of the very difficult 1SU4 → 1IWO transition (Weiss and Levitt, 2009) assuming that the intermediate structure 1VFP corresponds to a necessary pass-point in the transition. So, we decoupled our sampling methods to find the transition pathways that better accommodate this experimental information.

With the same Hamiltonian defined by the end states (A, B; see Methodological Approach), the Maxwell Demon MD is now first referred to the experimental intermediate structure (1VFP) and then switched to the target structure (when sampling of intermediate structure is converged). With this protocol, we selected the path closer to the putative intermediate state I. Figure 5 shows the obtained results. A clear improvement on the sampling of I is observed, showing that while the potential

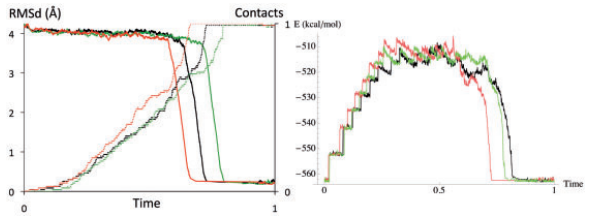


Fig. 4. Typical RMSd to target and potential energy profiles. Energy is presented in relative units. Different (three for simplicity) trajectories lead to similar but not identical results

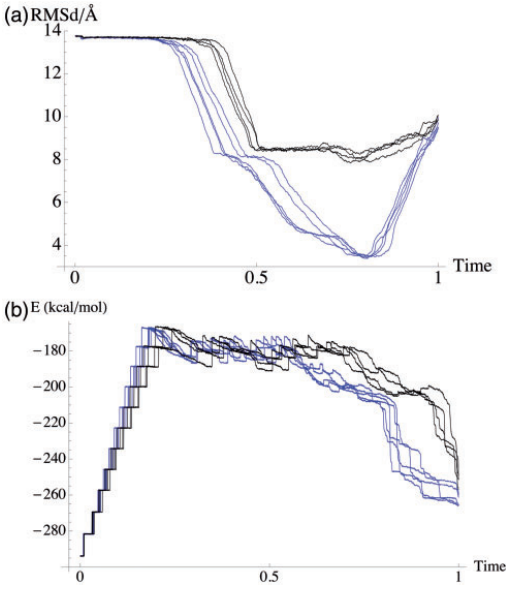


Fig. 5. 1SU4 → 1IWO transition including (blue) or not (black) information on the existence of an intermediate 1VFP Transition is repeated five times to guarantee reproducibility. (**a**) Profiles of RMSd with respect to experimental intermediate structure. (**b**) Potential energy evolution along the trajectory

energy surface is defined considering only A and B, the transition derived is fully compatible with the presence of the putative intermediate I.

Note that the new bias introduced in the trajectory by the experimental restrain does not impact dramatically on the energy, and it even improves energy relaxation toward the final state (Fig. 5).

4 CONCLUSION

We present GOdMD, a new method to trace conformational transition at the CG level. The method uses the fast CG dMD algorithm coupled to two biasing methods: (i) metadynamics adapted to follow proteins easiest deformation pattern, and (ii) a Metropolis-based Maxwell Demon algorithm. The method is fast and, contrary to many other approaches, always finds a smooth transition path when tested in a large dataset of transitions. Obtained trajectories maintain the covalent structure of the protein avoiding also unfavourable contacts, sampling intermediate structures with 'protein-like' properties, a clear advantage with respect to usual morphing schemes. The method is extremely flexible and can be easily adapted to introduce the effect of external perturbations in the transition or to bias the transition using experimental information. Limitations of the method are likely to be coupled to its CG nature that precludes the derivation of atomistic information on the transition. The approach is then expected to be especially powerful to obtain rough transition pathways that will be further refined by atomistic MD-based algorithms. A computer program implementing our method is publicly available as a web server at mmb.irbbarcelona.org/GOdMD.

Funding: Ministerio de Economia y Competitividad. Spain, the European Research Council (ERC-Advanced Grant (BIO2012-32868 to M.O.); the Instituto Nacional de Bioinformatica (INB); the Consolider E-Science Project (to M.O.); Framework VII Scalalife Project (to M.O.); Fundación Marcelino Botin (to M.O.). P.S. is a 'La Caixa fellow' and M.O. is an ICREA Academia Researcher.

Conflict of Interest: none declared.

REFERENCES

Bahar,I. and Rader,A.J. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, **15**, 586–592.
Bahar,I. *et al.* (2010) Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.*, **39**, 23–42.
Ban,D. *et al.* (2011) Kinetics of conformational sampling in ubiquitin. *Angew. Chem. Int. Ed. Engl.*, **50**, 11437–11440.
Beckstein,O. *et al.* (2009) Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open↔ closed transitions. *J. Mol. Biol.*, **394**, 160–176.
Best,R.B. *et al.* (2005) Slow protein conformational dynamics from multiple experimental structures: the helix/sheet transition of arc repressor. *Structure*, **13**, 1755–1763.
Bolhuis,P.G. *et al.* (2002) Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, **53**, 291–318.
Das,P. *et al.* (2006) Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. USA*, **103**, 9885–9890.
Delarue,M. and Sanejouand,Y.H. (2002) Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J. Mol. Biol.*, **320**, 1011–1024.
Dill,K.A. and Chan,H.S. (1997) From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.*, **4**, 10–19.
Ding,F. and Dokholyan,N.V. (2008) Dynamical roles of metal ions and the disulfide bond in Cu, Zn superoxide dismutase folding and aggregation. *Proc. Natl. Acad. Sci. USA*, **105**, 19696–19701.
Ding,F. *et al.* (2005) Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophys. J.*, **88**, 147–155.
Ding,F. *et al.* (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.
Ding,F. *et al.* (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat. Methods*, **9**, 603–608.
Dobbins,S.E. *et al.* (2008) Insights into protein flexibility: the relationship between normal modes and conformational change upon protein-protein docking. *Proc. Natl. Acad. Sci. USA*, **105**, 10390–10395.
Dror,R.O. *et al.* (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.*, **41**, 429–452.
Eisenmesser,E.Z. *et al.* (2002) Enzyme dynamics during catalysis. *Science*, **295**, 1520–1523.
Emperador,A. *et al.* (2008a) Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophys. J.*, **95**, 2127–2138.
Emperador,A. *et al.* (2008b) United-atom discrete molecular dynamics of proteins using physics-based potentials. *J. Chem. Theory Comput.*, **4**, 2001–2010.
Emperador,A. *et al.* (2010) Protein flexibility from discrete molecular dynamics simulations using quasi-physical potentials. *Proteins*, **78**, 83–94.
Emperador,A. *et al.* (2013) Efficient relaxation of protein–protein interfaces by discrete molecular dynamics simulations. *J. Chem. Theory Comput.*, **9**, 1222–1229.
Falke,J.J. (2002) Enzymology. a moving story. *Science*, **295**, 1480–1481.
Fenwick,R.B. *et al.* (2011) Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. *J. Am. Chem. Soc.*, **133**, 10336–10339.
Flores,S. *et al.* (2006) The database of macromolecular motions: new features added at the decade mark. *Nucleic Acids Res.*, **34**, D296–D301.

Franklin,J. *et al.* (2007) MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. *Nucleic Acids Res.*, **35**, W477–W482.
Gerstein,M. and Krebs,W. (1998) A database of macromolecular motions. *Nucleic Acids Res.*, **26**, 4280–4290.
Gherghe,C.M. *et al.* (2009) Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. *J. Am. Chem. Soc.*, **131**, 2541–2546.
Henzler-Wildman,K.A. *et al.* (2007) Intrinsic motions along an enzymatic reaction trajectory. *Nature*, **450**, 838–844.
Karplus,M. and Kuriyan,J. (2005) Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. USA*, **102**, 6679–6685.
Kern,D. and Zuiderweg,E.R. (2003) The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.*, **13**, 748–757.
Kim,M.K. *et al.* (2002) Efficient generation of feasible pathways for protein conformational transitions. *Biophys. J.*, **83**, 1620–1630.
Krebs,W.G. and Gerstein,M.B. (2000) The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res.*, **28**, 1665–1675.
Kubitzki,M.B. and de Groot,B.L. (2008) The atomistic mechanism of conformational transition in adenylate kinase: a TEE-REX molecular dynamics study. *Structure*, **16**, 1175–1182.
Laio,A. and Parrinello,M. (2002) Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA*, **99**, 12562–12566.
Leo-Macias,A. *et al.* (2005) An analysis of core deformations in protein superfamilies. *Biophys. J.*, **88**, 1291–1299.
Leone,V. *et al.* (2010) Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.*, **20**, 148–154.
Lezon,T.R. *et al.* (2009) Global Motions of the Nuclear Pore Complex: Insights from Elastic Network Models. *PLoS Comput Biol*, **5**, e1000496.
Lindahl,E. *et al.* (2006) NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Res.*, **34**, W52–W56.
Lindorff-Larsen,K. *et al.* (2005) Simultaneous determination of protein structure and dynamics. *Nature*, **433**, 128–132.
Lindorff-Larsen,K. *et al.* (2011) How Fast-Folding Proteins Fold. *Science*, **334**, 517–520.
Liphardt,J. *et al.* (2002) Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. *Science*, **296**, 1832–1835.
Lopez-Blanco,J.R. *et al.* (2011) iMod: multipurpose normal mode analysis in internal coordinates. *Bioinformatics*, **27**, 2843–2850.
Maragakis,P. and Karplus,M. (2005) Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.*, **352**, 807–822.
Marrink,S. *et al.* (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, **111**, 7812–7824.
McCammon,J.A. *et al.* (1977) Dynamics of folded proteins. *Nature*, **267**, 585–590.
Mendez,R. and Bastolla,U. (2010) Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. *Phys. Rev. Lett.*, **104**, 228103.
Meyer,T. *et al.* (2010) MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure*, **18**, 1399–1409.
Micheletti,C. (2013) Comparing proteins by their internal dynamics: Exploring structure-function relationships beyond static structural alignments. *Phys. Life Rev.*, **10**, 1–26.
Miyashita,O. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl. Acad. Sci. USA*, **100**, 12570–12575.
Okazaki,K.I. *et al.* (2006) Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, **103**, 11844–11849.
Orellana,L. *et al.* (2010) Approaching elastic network models to molecular dynamics flexibility. *J. Chem. Theory Comput.*, **6**, 2910–2923.
Orozco,M. *et al.* (2011) Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings. *Adv. Protein Chem. Struct. Biol.*, **85**, 183–215.
Perilla,J.R. *et al.* (2010) Computing ensembles of transitions from stable states: dynamic importance sampling. *J. Comput. Chem.*, **32**, 196–209.
Portella,G. and Orozco,M. (2010) Multiple routes to characterize the folding of a small DNA hairpin. *Angew. Chem. Int. Ed. Engl.*, **49**, 7673–7676.
Proctor,E.A. *et al.* (2011) Discrete molecular dynamics. *WIREs Comput. Mol. Sci.*, **1**, 80–92.
Proctor,E.A. *et al.* (2012) Discrete molecular dynamics distinguishes nativelike binding poses from decoys in difficult targets. *Biophys. J.*, **102**, 144–151.
Rueda,M. *et al.* (2004) Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations? *Biophys. J.*, **87**, 800–811.
Schlitter,J. (1994) Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graph.*, **12**, 84–89.
Sfriso,P. *et al.* (2012) Finding conformational transition pathways from discrete molecular dynamics simulations. *J. Chem. Theory Comput.*, **8**, 4707–4718.
Stein,A. *et al.* (2011) A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. *Structure*, **19**, 881–889.
Taketomi,H. *et al.* (1975) Studies on protein folding, unfolding and fluctuations by computer simulation. *Int. J. Pept. Protein Res.*, **7**, 445–459.
Tozzini,V. (2005) Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.*, **15**, 144–150.
Ueda,Y. *et al.* (1978) Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme. *Biopolymers*, **17**, 1531–1548.
Urbanc,B. *et al.* (2004) In silico study of amyloid beta-protein folding and oligomerization. *Proc. Natl. Acad. Sci. USA*, **101**, 17345–17350.
Urbanc,B. *et al.* (2010) Elucidation of amyloid β-protein oligomerization mechanisms: discrete molecular dynamics study. *J. Am. Chem. Soc.*, **132**, 4266–4280.
Velazquez-Muriel,J.A. *et al.* (2009) Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Struct. Biol.*, **9**, 6.
Weiss,D.R. and Levitt,M. (2009) Can morphing methods predict intermediate structures? . *J. Mol. Biol.*, **385**, 665–674.
Whitford,P.C. *et al.* (2007) Conformational transitions of adenylate kinase: switching by cracking. *J. Mol. Biol.*, **366**, 1661–1671.
Wiederstein,M. and Sippl,M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.*, **35**, W407–W410.
Yang,Z. *et al.* (2009) Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PLoS Comput. Biol.*, **5**, e1000360.
Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
Zhou,Y. and Karplus,M. (1997) Folding thermodynamics of a model three-helix-bundle protein. *Proc. Natl. Acad. Sci. USA*, **94**, 14429–14432.

Chapter 5: Predicting protein conformers

So far, we presented a systematic way of expanding the conformational landscape of proteins using two known structures. Here we designed an algorithm that takes advantage of the computational efficiency of GOdMD to predict transition path using coevolution information. The sampling engine, developed in our group, follows transitions path that maximize the coincidence of three-dimensional residue contacts with sequence coevolutionary information.

Coevolution information, or more specifically, coevolution contacts are sequence positions that show correlated mutations over the history of the protein. They can be inferred from a multiple sequence alignment of at least 2000 sequences, which is a requirement of our prediction protocol. Residue coevolution has been successfully applied to fold proteins (*244*, *245*), with remarkable accuracy in trans-membrane proteins (*246*). Extending the coevolution analysis to two binding proteins led to the identification of the complex geometry in docking experiments (*247*). An extend review of coevolutionary contacts applications is found elsewhere (*248*). Also, and very recently, coevolution contacts were used to characterize protein dynamics (*249*, *250*) and as a proof of concept in structural change predictions (*236*).

Based on the assumption that coevolving pairs of residues should be brought close in space in protein dynamics, we developed a dMD-based method to identify functional-relevant alternative conformational states. We benchmarked our method with over one hundred known structural transitions in the PDB. The method explored the power of coevolution analysis, as well as provided methodological advances to supplement biased molecular dynamics and coarse-grained models.

Title: **Residues coevolution guides the systematic identification of alternative functional conformations in proteins**

Authors: Pedro Sfriso*, Miquel Duran-Frigola*, Roberto Mosca, Agustí Emperador, Patrick Aloy, and Modesto Orozco

Stage: In press

Journal: Structure

Type: Research Article

Supplementary Material:

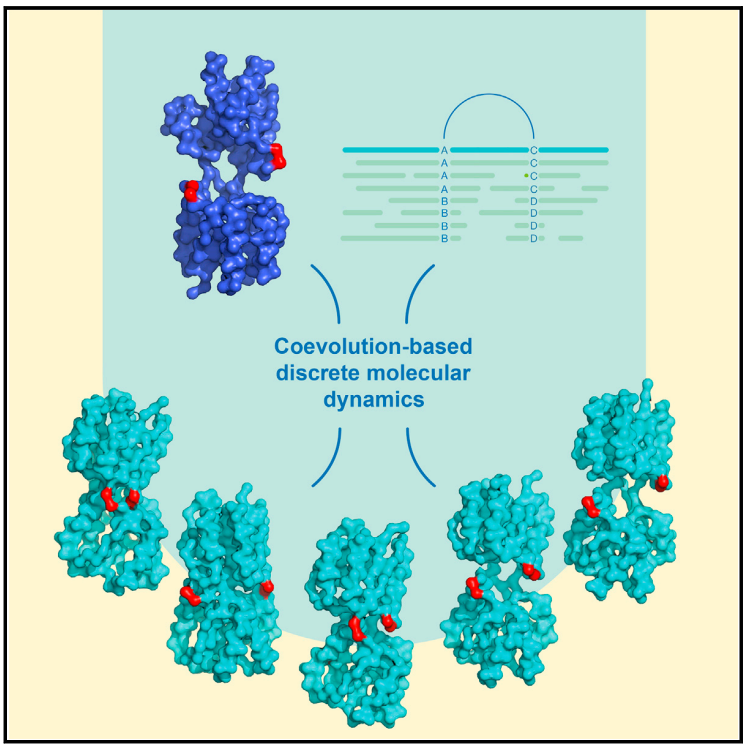
Author Contribution: PS was the main responsible all the work together with MDF. PS developed the simulation and sampling method and ran the simulations. PS contributed to the writing the paper.

Article

Structure

Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins

Graphical Abstract



Authors

Pedro Sfriso, Miquel Duran-Frigola, Roberto Mosca, Agustí Emperador, Patrick Aloy, Modesto Orozco

Correspondence

modesto.orozco@irbbarcelona.org (M.O.), patrick.aloy@irbbarcelona.org (P.A.)

In Brief

Protein flexibility is as important as structure to determine biological function. Sfriso et al. present a new approach, based on discrete molecular dynamics simulations guided by coevolutionary information, for the systematic identification of functional conformations in proteins. The strategy is able to capture alternative conformational states of varying complexity.

Highlights

- Automated prediction of alternative conformations in proteins
- Systematically explores the correlation between coevolution and dynamics
- Emphasizes the need for improving coevolution contact detection methods

Sfriso et al., 2016, Structure *24*, 1–11
January 5, 2016
http://dx.doi.org/10.1016/j.str.2015.10.025

CellPress

Please cite this article in press as: Sfriso et al., Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins, Structure (2016), http://dx.doi.org/10.1016/j.str.2015.10.025

Structure
Article

Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins

Pedro Sfriso,[1,2,5] Miquel Duran-Frigola,[1,2,5] Roberto Mosca,[1,2] Agustí Emperador,[1,2] Patrick Aloy,[1,2,3,*] and Modesto Orozco[1,2,4,*]
[1]Institute for Research in Biomedicine (IRB Barcelona), C/Baldiri Reixac 10, 08028 Barcelona, Spain
[2]Joint BSC-IRB Research Program in Computational Biology, C/Baldiri Reixac 10, 08028 Barcelona, Spain
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluis Companys 23, 08011 Barcelona, Spain
[4]Department of Biochemistry and Molecular Biology, University of Barcelona, Av. Diagonal 647, 08028 Barcelona, Spain
[5]Co-first author
*Correspondence: modesto.orozco@irbbarcelona.org (M.O.), patrick.aloy@irbbarcelona.org (P.A.)
http://dx.doi.org/10.1016/j.str.2015.10.025

SUMMARY

We present here a new approach for the systematic identification of functionally relevant conformations in proteins. Our fully automated pipeline, based on discrete molecular dynamics enriched with coevolutionary information, is able to capture alternative conformational states in 76% of the proteins studied, providing key atomic details for understanding their function and mechanism of action. We also demonstrate that, given its sampling speed, our method is well suited to explore structural transitions in a high-throughput manner, and can be used to determine functional conformational transitions at the entire proteome level.

INTRODUCTION

Proteins are not rigid structures but flexible and dynamic entities, which adapt their conformations to respond to cellular stimuli, perform mechanical work, catalyze biochemical reactions, or interact with other macromolecules (Eisenmesser et al., 2002; Henzler-Wildman and Kern, 2007; Stein et al., 2011). There is a bulk of evidence demonstrating that flexibility is as important as structure in defining the function of proteins (Falke, 2002; Henzler-Wildman et al., 2007; Micheletti, 2013; Orozco, 2014), and that evolution has made a big effort to maintain and refine the functionally relevant conformational space of proteins (Leo-Macias et al., 2005; Stein et al., 2011; Velazquez-Muriel et al., 2009).

Often protein flexibility arises from near-equilibrium dynamics, i.e. from the activation of essential deformation modes of the native structure (Bahar et al., 2010; Das et al., 2014). In these simple cases, alternative conformations are located in a pseudo-harmonic free-energy funnel centered at the equilibrium state, and can be sampled from short-timescale molecular dynamics (MD) simulations (McCammon et al., 1977), or even from simple coarse-grained elastic network model calculations (Kim et al., 2002; Yang et al., 2007). However, there are also more complex instances whereby proteins have to undergo large conformational transitions to perform their biological function. These distant conformers are very difficult to predict from theoretical methods designed to sample around equilibrium geometries of known structures. Pure force atomistic MD simulations are an obvious alternative in these cases (Dror et al., 2012; Shimamura et al., 2010), but even with specific-purpose computers, the accessible timescale for MD moves in the submicrosecond to millisecond range (Shaw et al., 2010), still far from the timescale of many functionally relevant transitions. Coupling of MD simulations with biasing techniques (Elber and West, 2010; Elber, 2005, 2007; Laio and Parrinello, 2002; Perilla et al., 2010; Sfriso et al., 2012) permits exploration of conformational transitions that happen on timescales slightly above those accessible from unbiased MD. These techniques are not only very CPU-demanding, but also require some previous knowledge on the transition pathway, which limits their applicability for predicting unknown protein conformations or determining new conformational pathways.

It is clear that, while waiting for more accurate force fields, better biasing techniques, and faster computers, the only way to explore the vast conformational space is to incorporate experimental restraints into the theoretical calculations (Chen and Hub, 2014; Seeliger and de Groot, 2010; van den Bedem et al., 2013). Thus, structural data derived from electron microscopy, nuclear magnetic resonance, or X-ray crystallography have been used to help theoretical methods to trace large transitions able to capture different conformational states, typically by defining the start and end conformations of the protein (Beckstein et al., 2009; Sfriso et al., 2012; Weiss and Levitt, 2009; Whitford et al., 2007). Unfortunately, this paradigm of integration of experiment and simulation is applicable only when at least one distant alternative conformation of the target protein is known. In other words, we have powerful methods to explore structural states within transition paths between two known conformers, but such methods cannot identify alternative functionally relevant conformations.

Coevolutionary data have been used as a source of indirect structural information on proteins allowing, in very favorable cases, the determination of the folded state (Hopf et al., 2012, 2015; Jones et al., 2015; Marks et al., 2011, 2012; Michel

Please cite this article in press as: Sfriso et al., Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins, Structure (2016), http://dx.doi.org/10.1016/j.str.2015.10.025

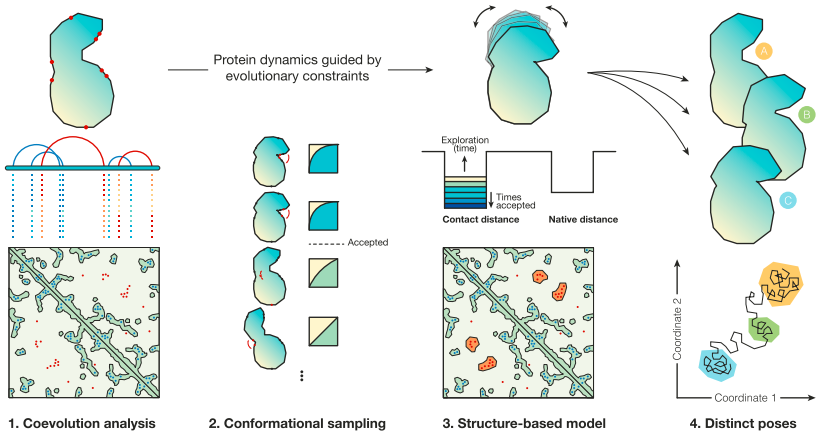


Figure 1. Method Summary

The protocol uses (1) raw coevolution DCA scores to (2) test the accessibility of each residue pair in the structure by means of an initial conformational sampling. Individual trajectories are accepted when they show better coincidence with coevolution information than a threshold (area under the ROC curve, see Experimental Procedures). If consistency is observed between coevolution data and the conformational sampling, we (3) incorporate the corresponding pairs of residues into SBMs. Coevolution pairs are reflected in the models by favorable energy interactions, exploring the conformational landscape accordingly. Implicitly, this approach filters noise in the DCA signal, and reveals the protein ensemble encoded by coevolution. Finally, we (4) select distinct conformations from the dMD simulations to provide a small set of structures that is representative of the conformational landscape.

et al., 2014; Morcos et al., 2011), and the trace of simple open-to-closed transitions (Morcos et al., 2013). Here, we further explore the power of the coevolutionary signal to guide theoretical methods in the search for conformational ensembles and alternative functionally relevant conformations.

We present an automated protocol whereby coevolution contacts are filtered and introduced as ensemble restraints in coarse-grained discrete molecular dynamics (dMD) simulations, which are able to detect alternative functionally relevant conformations. We validate the predictive power of the method on an exhaustive set of alternative structural states extracted from the PDB. We found that in 76% of proteins studied the protocol is capable of finding an alternative conformer. Finally, we assess the general applicability of our method to explore conformational transitions of varying complexity, including a prediction of serine/threonine protein kinase conformers. Predicted conformers can be found at mmb.pcb.ub.es/CBDMD/.

RESULTS AND DISCUSSION

The protocol developed, as outlined in Figure 1, is based on four consecutive steps. First, we performed direct coupling analysis (DCA) (Weigt et al., 2009) on a multiple sequence alignment, selecting those coevolving pairs of residues that are not in contact in the native structure, and which might thus be informative of alternative protein conformations. In a second step, we cleaned the DCA output to remove uninformative or impossible contact pairs. To this end, we used dMD (Proctor et al., 2011; Sfriso et al., 2015) to bring coevolution pairs close in space (one independent dMD simulation for each pair), up-ranking viable trajectories leading to conformations that are coherent with the rest of the coevolution map (see Figure S1). In a third step, after selecting the most informative coevolution pairs, we built structure-based models (SBMs) (Taketomi et al., 1988; Tozzini, 2005; Ueda et al., 1978; Whitford et al., 2007) to perform coevolution-biased discrete molecular dynamics (cb-dMD) simulations. Finally, we clustered and analyzed the trajectories to generate an ensemble of representative conformers, which are expected to represent the functionally relevant conformational landscape of the protein (see the Experimental Procedures for further details).

Sufficient Coevolutionary Information Enables Systematic Detection of Alternative Conformers

To validate the method, we explored its ability to detect known alternative conformations in a set of proteins with more than one structure available in the PDB (Berman et al., 2000), ensuring sufficient protein coverage and coevolutionary signal by filtering out sequences with less than 2,000 members in the alignment (Figure S2). A robust non-trivial validation test was defined by filtering out pairs of structures separated by less than 3 Å in root-mean-square deviation (RMSD), since these limited conformational transitions could be captured by standard equilibrium-dynamics methodologies. Redundant proteins (sequence identity >70%) were also discarded. The resulting validation set contained 105 proteins. We labeled two source structures (A/B) per protein, defining a total of 210 transitions to be determined (when more than two conformers were found in databases, we selected the two with best sequence coverage, provided they were at a distance >3 Å) (Figure 2A). We ran our method on the source structures, and after clustering each trajectory we retrieved ten representative conformers. In 13 of the 105 proteins, none of the predicted conformers satisfied any exclusive coevolution contact, and we excluded them from the validation set. These 13 proteins corresponded mainly to closed-to-open transitions not suitable for reproduction by our coevolution-based method, since the sampling engine requires unique coevolved contacts in the alternative conformers. The final validation set contained 92 proteins corresponding to 140 source structures (195 transitions) (Table S1 and Figure 2A), each of them yielding an ensemble of ten conformers with coevolving residues forming new contacts. Next, we checked whether these conformers approached the experimental ones by measuring the RMSD, and also compared the overlap of the expected transition (between two known end points A and B) with the transition from the source structure to a predicted conformer. For benchmark purposes, we computed an experimental p value of the overlap obtained with our protocol using a background distribution of overlaps (obtained from a converged equilibrium simulation; see the Experimental Procedures). Cases with a high overlap ($p < 0.05$) between predicted and expected conformers were considered to be successful.

Under the criteria explained above, we recalled at least one known alternative conformation for 59% of the source structures (Figure 2C), consistently approaching the target state (Figure 2E), mostly detecting one conformer per case (Figure 2F), and being on average 4.5 of the 10 predicted poses relevant (Figure 2G). Overall, we identified alternative conformers in 70 of the 92 proteins considered, leading to a success rate of 76% (Figure 2D). Therefore, when B-to-A and A-to-B transitions were both studied it was very likely that we identified alternative conformers, particularly in open-closed motions (83%). Worst performances were achieved for domain rotation movements (69% success), wherein the formation of new coevolution contacts is less concerted or barely existent.

The selection step enriches in informative long-range DCA contacts (Figure 3A), which is a key step in our protocol. Typically, our method accepts ~10% of original DCA contacts, those with best area under the curve (AUC) score (see Figure S1). On average, we added 1,187 coevolution-based wells in the dMD energy potential for each simulation, which represents about 19% of the total potential energy interaction. The average number of models used to derive this SBM is 83, which are in turn used to bias the simulation (see Sampling strategy in the Experimental Procedures). The specific details for each system can be found in Table S2.

The importance of the selection step is tightly related to the type of movements analyzed above. To obtain further insight into this issue, we investigated the impact of the number of sequences on the quality of the different motions. We randomly removed sequences from the alignment and re-ran our protocol with 1,000, 2,000, 3,000, 4,000, 5,000 and 10,000 sequences for 20 cases, spanning open-closed, rotation, rotation-closed, concerted, and miscellanea of complex motions. Figure 3B shows that coevolutionary signal relevant to rotations steadily decreases as sequences are removed from the alignments, while open-closed transitions are less sensitive to the alignment size, suggesting that the depth of coevolution information required depends on the characteristics of the movement and the available conformation in the PDB. Open-closed transitions evince exclusive contacts (71 ± 50) easier than, e.g., rotations (29 ± 30), and exclusive contacts in target conformation in turn differentiate successful cases from unsuccessful ones (Wilcoxon's p value 3.3×10^{-5}).

Not surprisingly, to obtain successful simulations in these cases larger sequence alignments were required: often we needed ~16,000 sequences to reproduce rotations. The average number of sequences in the successful open-closed cases was only of ~8,000, and decreased to the pre-set minimum of 2,000.

Unique Capacity to Identify Varied, Non-trivial Conformations

In the validation set, of the 70 successful cases 15 underwent open-closed movements, 11 rotations, 15 rotation-closed motions, and 14 concerted motions, and the remaining 16 a miscellanea of complex transitions. Coevolution data are thus applicable to many scenarios. This trait is better depicted in Figure 4, which displays transitions of varied extent and complexity, from helix translocations to domain-domain rearrangements. To assess the relevance of the predictions we compared our results with ensembles generated with other control methods (Figure 2H). In our hands, coil, equilibrium, and normal-mode analysis (NMA)-based methods were not able to capture such a spectrum of movements. These controls demonstrate the unique capacity of our approach to identify non-trivial alternative conformations, which reach beyond equilibrium fluctuations and are not accessible by essential deformation movements (as defined, e.g., through NMA). As an additional control, we implemented a direct coevolution-based SBM that simply uses DCA contacts as energy minima (Morcos et al., 2013). Compared with ours, this direct technique showed less accurate results (Figures 2H and 2I), which advocates for the relevance of the filtering step included in our protocol.

Following the observation above, we studied in more detail the contribution of the aforementioned filtering of coevolution pairs (Figure 1; pulling trajectories). In this key step, only coevolved pairs that lead to coherent deformations are retained. We observed that this filter was not critical when abundant sequences were available. For instance, we were able to collect 14,893 sequences for the D-ribose binding protein (PDB: 2DRI, chain A), yielding strong evolutionary signal. In this case, both our method and the direct incorporation of the coevolution map were able to reproduce the large closure from the open conformation (PDB: 1BA2 A), and even to detect other transient states (Morcos et al., 2013) (Figure 5A). However, a similar conformational transition turned out to be more challenging for the direct method when fewer homologs could be aligned, as in the case of 5-enolpyruvylshikimate 3-phosphate synthase (2,271 sequences) (Figure 5B). Here, the unfiltered, direct inclusion of coevolving pairs was not able to produce any relevant movement, due to noise in the coevolution map. On the contrary, our method traced the long-range transition from the open (PDB: 1EPS A) to the closed state (PDB: 2AAY A) without erratic exploration of the conformational space. These results are in good agreement with our observations in Figure 3B, where a good proportion of the pairs relevant to protein dynamics are still retrieved with a relatively small number of sequences aligned. In our experience, few high-quality coevolved pairs are necessary to robustly guide protein dynamics, making the detection of these constraints decisive, and suggesting that coevolution-driven dynamics for mid-size families is feasible if coevolution data are carefully filtered.

Biased Structure-Based Models Yield Smooth Multi-State Transitions

A second key step for the success of our approach is the compilation of structures in the SBM (Figure 1). When more than one conformation is captured in the coevolution footprint, or if two (or more) domains move concertedly, extracting information from the coevolution contacts is far from trivial and, accordingly, predicting functional transitions is difficult. An example of the former is the conformational transition undergone by *Escherichia coli* adenylate kinase (PDB: 4AKE A to 1AKE A). This kinase performs a coordinated two-domain closing motion (Figure 5C), whereby LID and AMP-binding domains approximate to complete the shift from an apo to a holo state. This two-domain transition is nicely reproduced by our method using only one source structure, with no additional information on the target conformer. Capturing the two parts of the motion in the pulling trajectories, a

Please cite this article in press as: Sfriso et al., Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins, Structure (2016), http://dx.doi.org/10.1016/j.str.2015.10.025

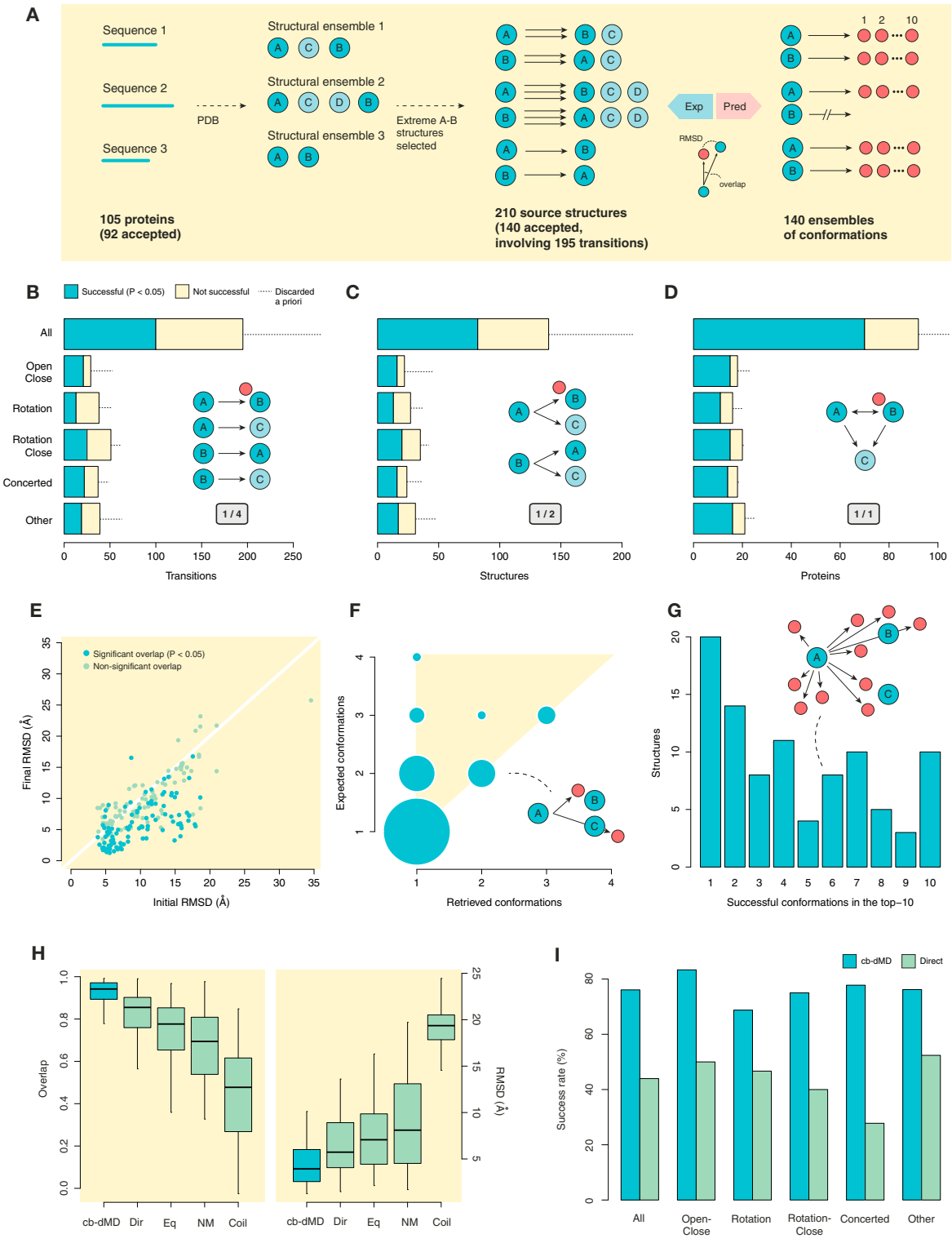


Figure 2. Method Validation

(A) Flowchart of the validation set selection. From the PDB, we kept ensembles with at least 99% sequence coverage in one of the structures. "Accepted" refers to cases not discarded a priori.

(legend continued on next page)

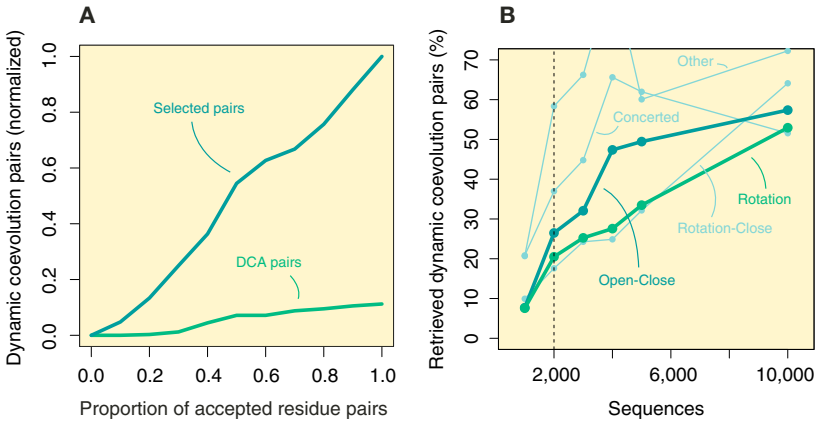


Figure 3. Sequence Number Effect and Selection of Coevolution Pairs

(A) Enrichment in dynamic coevolution pairs among the selected list, compared with the DCA list after removing native contact pairs. Dynamic coevolution pairs are those pairs of residues that are not in contact in the source structure but are proximal in the target structure. Lines in the plot are the average of the benchmarked trajectories: in general, thus, the filtering step selects pairs that will be useful to guide the molecular simulation toward the target state.

(B) Coevolutionary signal kept depending on the number of sequences and the type of motion: for simpler open-close transitions fewer sequences are needed compared with rotation motions, where only few contacts per conformer are exclusive. Sequences were randomly removed from initial alignments. Retrieved dynamic pairs correspond to pairs that were accepted in the pulling trajectories.

unique feature of the approach, guides the transition even with a reduced number of sequences aligned (2,034).

Some proteins elicit yet more complex movements following pathways through multiple states. If functionally relevant, these states should also be preserved by evolution, and thus explored and connected by our method. One clear example (Figure 5D) is the long-chain fatty acid-coenzyme A ligase (PDB: 1ULT A) motion, with two alternative structures available, namely PDB: 1ULT B and 1V26 A. Along the trajectories, we spontaneously sampled configurations similar to all known alternative conformers, suggesting that our protocol is able to span the conformational landscape associated with the mechanism of catalysis (Hisanaga et al., 2004). It is worth noting that, when using the direct DCA approach (Morcos et al., 2013), most of the time the trajectory samples a compact conformation that does not resemble any of the known structures for this system.

Conformer Prediction Facilitates Mechanistic Interpretation

Finally, we propose predicted conformers for the PAS domain of the human serine/threonine protein kinase (PASK). Protein kinases are important drug targets, but structure-based drug design is often impeded by their intrinsic flexibility (Engh and Bossemeyer, 2002). Fortunately, kinases are large families with many sequences available, making them a valuable example of application of our protocol. Departing from the initial structure (PDB: 3DLS A), we gathered 21,840 sequences and proposed ten new conformers. Figure 6 illustrates the process of selecting a discrete number of conformers from the trajectory. We project each trajectory into its two first components (Figure 6A), and use DBSCAN (Ester et al., 1996) to extract the most dense cluster of conformations, five in this case (Figure 6C). We repeat this for ten independent trajectories to ensure robustness. After discarding structurally similar conformers, we rank the predicted conformers. We represent in Figure 6B the departing structure together with the top-ranked conformer. The conformational transition is moderate (4.7 Å RMSD) and, interestingly, it approaches the ATP-binding site (blue sphere) with the proton acceptor site (green sphere) and the P loop (orange), responsible for the phosphate transfer. The conformational landscape thus proposes a coarse, yet illustrative, mechanism of action. Kinase conformers, besides providing mechanistic insights of the phosphorylation process, could be used to test the possibility of auto-phosphorylation either in monomers or dimers, and be applied in structure-based drug design to improve ligand docking or to spot transient druggable cavities.

(B) Distinct trajectories reproduced by our method. The bars count the number of transitions, and the blue shading quantifies the number of successful cases (a case was considered as such when at least one of the top ten conformations largely overlapped the expected transition [$p < 0.05$]). Red circle indicates successful predicted structure.

(C) Similarly, performance is evaluated when at least one alternative conformation is found, departing from a source PDB structure. In this case, the blue shading denotes that for 59% of the source structures we could identify at least one alternative conformer.

(D) Finally, results analyzed at the protein level, i.e. departing either from A or B structures, show a success rate of 79%. Success rate ($p < 0.05$) of cb-dMD compared with direct incorporation of DCA pairs.

(E) Initial distance of the source structure to the target one versus the distance after running the pipeline. Successful simulations are highlighted in dark blue (significant overlap).

(F) For the successful cases, the expected and retrieved alternative conformers. This sketch outlines a scenario where two of the two expected conformations are approached. The most common scenario (denoted with the bigger circle) is that of only one alternative conformer being approached.

(G) Number of retrieved known conformations. Most trajectories expect and find only one conformation.

(H) Our method (cb-dMD) is compared with an implementation that directly incorporates all DCA pairs (Morcos et al., 2013) (Dir) as energy minima, with a Go-like equilibrium simulation (Eq), with a normal-mode guided sampling (NM), and with a background random-coil polymer (Coil). Here, overlaps and RMSD along the trajectory are displayed to avoid the impact of the clustering step, which would penalize controls.

(I) Comparison of our method with the direct implementation of DCA pairs, for each motion type.

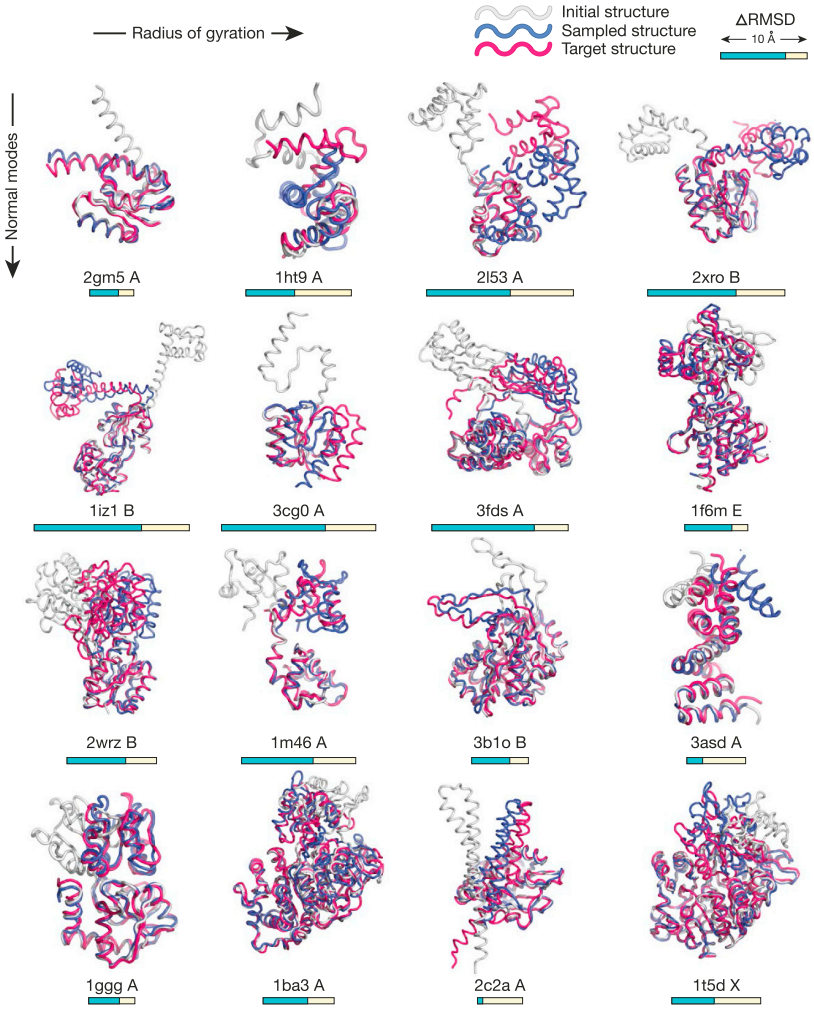


Figure 4. Representative Space of Captured Movements

Gray structures represent the departing structure, while pink structures correspond to an alternative conformation reported in the PDB. Blue structures show the closest predicted alternative conformer. We manually selected transitions in a range of overlaps with the normal modes of the initial structure (vertical axis), and the relative change of the radius of gyration (RG; horizontal axis). Therefore, the bottom-left area of the figure corresponds to large overlaps to the normal modes (>0.80) and compaction (ΔRG ≈ −10%–20%) of the structure. Note that low normal-mode and large ΔRG motions are particularly challenging for our protocol due to the scarcity of unique contacts in alternative structures. Length of the bar below each ensemble is proportional to the RMSD between the two experimental structures. The blue bar represents the proportion of this distance traveled in the simulation.

Concluding Remarks

Overall, after studying dozens of cases, we have confirmed that the echo of correlations in protein evolution is tightly related to dynamics constraints. By exploiting residue-residue coevolution, we have enhanced the sampling of protein conformations, which are currently impossible to explore systematically by experiments. Our protocol is applicable to cases of varying complexity, requiring as input only a multiple sequence alignment of at least 2,000 sequences and one 3D structure. We are able to detect alternative conformations if they show unique subsets of coevolved contacts, and in complex scenarios we can identify multiple states, and the paths leading from one to the other, giving mechanistic and functional insights into the way protein families operate.

To visit functionally relevant states, we have seen that large multiple alignments and filtering of coevolution data are still crucial. In particular, we found that the selection of coevolution contacts was key to enabling the exploration of alternative conformers in cases with few sequences that simpler methods (Morcos et al., 2013) cannot reproduce. Currently the PDB contains 270,380 structures, of which we were able to obtain plausible multiple sequence alignments for 65,349 (24.14%), suggesting a broad applicability of our method. These structures correspond to 8,813 unique proteins in 1,542 species, and span 1,051 (15.25%) of the Pfam domains represented in the PDB (Finn et al., 2014). We envisage that the applicability of coevolution-based dynamics will increase even further in the near future, given the explosive growth of sequence databases (Khafizov et al., 2014), and the massive deposition of structures arising from structural genomics initiatives (Khafizov et al., 2014).

To date, coevolution analysis has been mainly applied to de novo structure prediction and, as protein sequences continue to accumulate, there is debate on the usefulness of coevolution methods for other applications (Kamisetty et al., 2013; de Juan et al., 2013). Recently, evolutionary information was used to understand allosteric mechanisms (Halabi et al., 2009) and, along this line, our findings are yet further proof of the importance of coevolution analysis for the structural biology community, here as a source of information to predict functional conformers. Interestingly, we have found that coevolving pairs that are relevant to dynamics rank far below those that are useful for protein folding (Figure S3), advocating for further development of coevolution analysis methods, and confirming that coevolutionary pressure acts beyond the mere preservation of contacts in the native structure.

EXPERIMENTAL PROCEDURES

Multiple Sequence Alignments

We use HHblits (Remmert et al., 2012) to align multiple sequences from the clustered UniProt database (March 2013). The following options are used in addition to default settings: -diff inf, -mact 0.5, -n 5, -cov 75, and -maxfilt 500,000. We discard alignment sites corresponding to gaps in the query

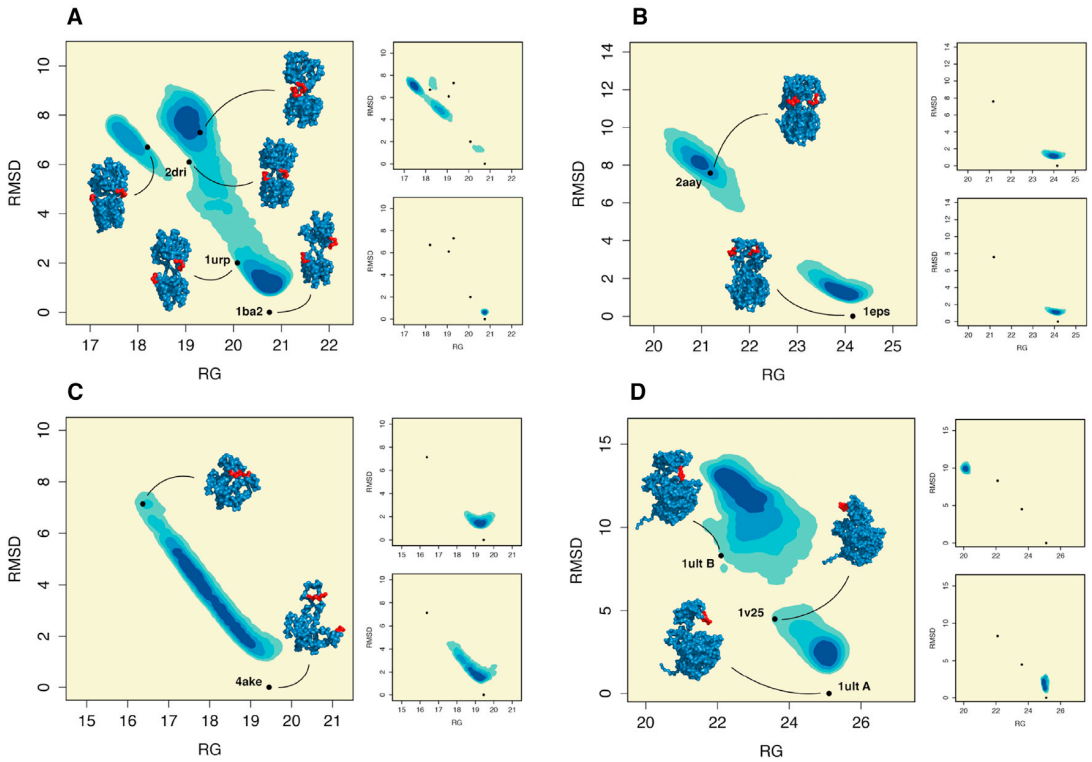


Figure 5. Detailed Case Examples

Bidimensional histograms of sampled structures; the x axis shows the radius of gyration (RG), and the y axis the distance to the initial structure. In each panel, the left plot represents our results, displaying relevant structures. The upper-right plot corresponds to simulations obtained upon the direct selection of top-ranked DCA pairs, without the filter based on pulling trajectories. The lower-right plot displays the trajectory obtained upon random coevolution maps (see Experimental Procedures), illustrating the relevance of the coevolution signal.

(A) Starting from 1BA2, we visited all known alternative configurations, including relevant intermediates, in good agreement with results obtained by others (Morcos et al., 2013).

(B) On the contrary, we could only observe the closing trajectory of PDB: 1EPS after filtering coevolution contacts, as uniquely done by our method.

(C) In the two-domain motion departing from 4AKE, the integration of multiple structures in the SBM was crucial to coordinate the transition.

(D) Departing from 1ULT A, we predicted a domain rotation, involving a rich conformational repertoire that was partially validated by structures deposited in the PDB.

sequence, in addition to those sites with more than 25% gaps along the alignment (Kamisetty et al., 2013).

Direct Coupling Analysis

To measure residue-residue coevolution, we use DCA with default parameters: x = 0:2; λ = 0:5 (Weigt et al., 2009). DCA outputs a direct information (DI) score per pair of residues, ranking evolutionary correlation. Only coevolution pairs at a sequence distance $\geq$5 are considered.

Selection of Coevolution Pairs

Given a DI-ranked list of coevolution pairs, we keep for further analysis only the first n pairs that maximize the Matthew's Correlation Coefficient (MCC) resulting from the prediction of contacts (<10 Å) in the initial structure. Given a list of n selected coevolution pairs, MCC is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(FP+FN)(TN+FP)(TN+FN)}}, \qquad \text{(Equation 1)}$$

where TP is the number of contacts in the selected list, while FP corresponds to the selected pairs that are not in contact, TN to the non-selected pairs that are not in contact, and FN to contacts that have not been selected. The intuitive interpretation of this step is that we extend to a larger number of DCA contacts (ordered by their DI score) while they are still informative about the initial structure.

Exploratory Conformational Sampling Based on Coevolution

Coevolution pairs that are far apart in the initial structure are used to guide an initial, exploratory conformational sampling. Concretely, we run a pulling trajectory for each distant coevolution pair i-j using dMD. According to the standard dMD algorithm, the protein Hamiltonian is defined as a series of flat square wells (in this case Go-like single well, see Figure S4 and Emperador et al., 2008a), and particles (Cα) move at constant velocity until a collision occurs, where momentum and energy conservation rules are enforced. The dMD algorithm avoids femtosecond-scale integration of Newton's equations of motion, dramatically improving computational efficiency (Emperador et al., 2008b; Orozco et al., 2011; Shirvanyants et al., 2012). To favor the formation of coevolutionary contacts (i-j) we bias the trajectory by inverting velocities of particles i and j every 100 simulation steps if i and j are not approaching each other, and keeping them unaltered otherwise (we consider that two residues are in contact if they are at less than 10 Å). We permissively maintain the trajectory for downstream analysis if i-j are in contact at some point of the trajectory.

Selection of Compelling Pulling Trajectories

From a functional viewpoint, of all the preliminary trajectories generated above, the most interesting ones are those that are in better agreement with the coevolutionary signature. To evaluate the coincidence between trajectories and coevolution maps, we check whether contacts spontaneously established along the pulling trajectory ($k - l$, where $k,l \neq i,j$ are indeed

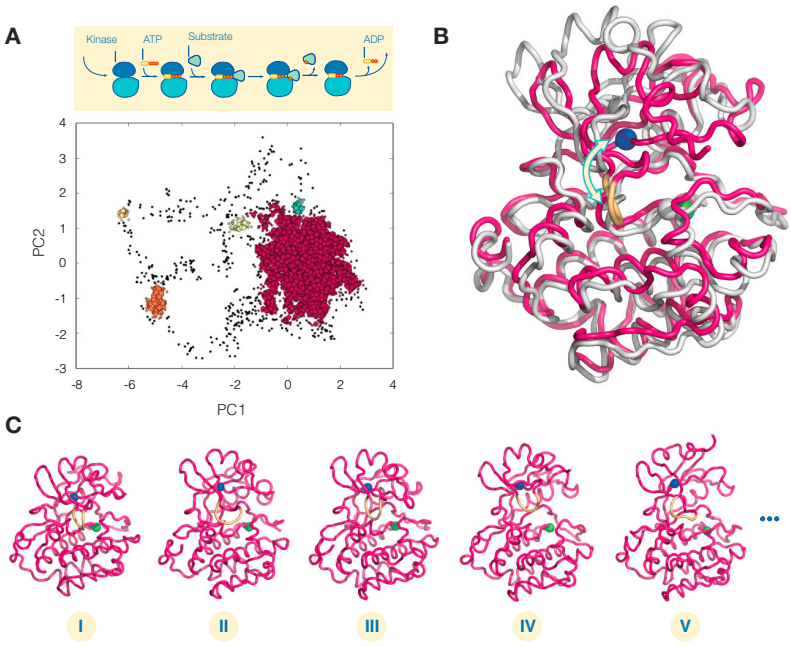


Figure 6. PASK Conformers

(A) 2D projection of a trajectory into its two first components (PC1, PC2) (see Experimental Procedures: Selection of Representative Structures). Colored dots indicate belonging to an automatically identified cluster, 5 in this example. A sketch of a generic kinase-substrate phosphorylation mechanism is depicted above the plot. (B) Best-ranked conformer (pink) together with the initial structure (white). Blue and green spheres represent ATP-binding site and proton acceptor site, respectively. (C) Best-ranked conformer (I), and the complete cluster ensemble for this trajectory. These are the proposed models that can be extended up to ten using independent replicas.

coevolution pairs). For this, we compute receiver-operating characteristic curves (ROC; sensitivity versus 1 − specificity) to quantify the agreement between conformations generated in the trajectory and the list of *n* coevolution pairs (we filter out those contacts at ≲6.5 Å to exclude trivial trajectories). In this framework, the ROC space is defined as follows:

$$Sensitivity = \frac{TP}{TP + FN}\,, \qquad \text{(Equation 2)}$$

$$Specificity = \frac{TN}{FP + TN}\,, \qquad \text{(Equation 3)}$$

where TP counts contacts generated along the trajectory that are in turn selected coevolution pairs, FN is the number of coevolution pairs that are not in contact along the trajectory, TN are the pairs that are not in contact during the trajectory and, accordingly, are not coevolution pairs, and FP are the pairs that are in contact but do not coevolve.

The AUC (area under the resulting ROC curve) provides a means to compare and rank the coherence between trajectories and the coevolutionary fingerprint. To select those trajectories that best coincide with the coevolution signal, we retain instances exceeding 1.5 of the interquartile range in the AUC distribution (see Figure S1). In these cases, we keep the last frame of each trajectory. The retained trajectories thus yield a set of seed conformations (*nmodels*) to be used in downstream analysis.

Structure-Based Modeling

Given the pulling trajectories described above, we build a multiple SBM that would enable a single dMD trajectory to explore the *nmodels* ensemble. To this aim, we shape particle-particle interactions to reflect the variability spanned by exploratory trajectories. Concretely, we take the original PDB and the last snapshot of the accepted pulling trajectories, and describe the potential energy interactions by means of double-well square potentials when a pair of particles i-j distance change across the ensemble of structures, and single-well square potential otherwise. The wells are centered at r_{ij}^{PDB} (the distance in the initial PDB structure), and, when needed, a second one centered at the coevolution interaction distance between residues $i-j$ (r_{ij}^{coev}; Equation 4). To set r_{ij}^{coev} we consider all *nmodels* $i-j$ distances coming from the accepted pulling trajectories. To increase the importance of short-distance contacts, which could be obscured by larger ones, we introduce a weight factor in the averaging (w_{ij}^k; Equation 5).

If *nmodels* is the number of structures used to build the SBM, the center of the coevolution well r_{ij}^{coev} is

$$r_{ij}^{coev} = \sum_{k=1}^{nmodels}\left(\frac{w_{ij}^k}{\sum_{k=1}^{nmodels} w_{ij}^k}\right) r_{ij}^k\,, \qquad \text{(Equation 4)}$$

$$w_{ij}^k = \frac{1}{\left(r_{ij}^k\right)^3 - (2R_{HC})^3}\,, \qquad \text{(Equation 5)}$$

where r_{ij}^k is the residues $i-j$ distance in the *k*th model, and the hard-core radius of the particles R_{HC} is 2 Å.

For wells representing the initial PDB distance, energy depth is $E(r_{ij}^{PDB}) = -0.30$ kcal/mol, a value that was adjusted to keep stable conformations for proteins at 300 K. To favor the robust coevolutionary signal, we deepen the associated depth by a factor ε = 1.05 every time the coevolution interaction coincides with the internal distance in a given model k ($r_{ij}^k \approx r_{ij}^{coev}$). So, if *N* is the number of models contributing to $i-j$ coevolution interaction, the energy associated to the well is

$$E\left(r_{ij}^{coev}\right) = -\varepsilon^N 0.30 \;\; kcal/mol\,. \qquad \text{(Equation 6)}$$

This discriminates coevolution contacts that are observed a few times with respect to the ones observed in several models.

Sampling Strategy

We adapted our GOdMD protocol (Sfriso et al., 2013) to explore the conformations captured in the multiple SBM. For this purpose, the biasing scheme was modified to visit multiple target states instead of reaching a single one. We construct the Γ function (Equation 7) to bias the trajectory toward the distinct protein poses captured in the multiple SBM. Γ reflects the variability in the SBM by summing up the internal distances of the accepted *nmodels*, r_{ij}^k being the internal distance of i-j pair in the *k*th model:

$$\Gamma = \sum_{k}^{nmodels}\sum_{i\,j} r_{ij}^k \delta_{ij}^k\,, \qquad \text{(Equation 7)}$$

where δ_{ij}^k is a Kronecker's variable that takes a value of 0 whenever a pair of particles are at r_{ij}^{coev} distance or shorter, and 1 otherwise. δ_{ij}^k is used to permanently eliminate the bias toward a given model *k* when all coevolution wells of this model were visited, which favors multiple-state exploration. If, despite the use of δ_{ij}^k, no progress was observed in the simulations (i.e. no novel coevolution-based wells were explored) we deactivate temporarily (5,000 time units [t.u.]) the biasing scheme to facilitate relaxation and escape from stationary points. Therefore, the sampling strategy consists in evaluating the Γ function at every Δ*t* = 10 t.u. of free dynamics to check whether the trajectory is sampling the conformational space revealed by coevolving residues. Accordingly, we accept with probability $p(\Gamma)$:

$$p(\Gamma) = \begin{cases} 1, & \Gamma_t < \Gamma_{t-\Delta t} \\ e^{-\beta(\Gamma_t - \Gamma_{t-\Delta t})^2}, & \Gamma_t \geq \Gamma_{t-\Delta t} \end{cases}\,, \qquad \text{(Equation 8)}$$

the latest Δ*t* of the trajectory, ensuring an exhaustive sampling of the accessible coevolution wells. β was introduced to keep the acceptation rate at suitable

values (50%–80%) (Sfriso et al., 2013). To further improve the computational efficiency, the GOdMD algorithm introduces an additional metadynamics procedure (Barducci et al., 2011; Laio and Parrinello, 2002), which penalizes any visited well by gradually reducing their depth (for further details see Sfriso et al., 2013).

Selection of Representative Structures

We obtain an estimate of the conformational space by running ten independent GOdMD trajectories. We then reduce the dimensionality of the conformational space sampled by projecting the snapshots to the two first principal components of the trajectory. Finally, we use DBSCAN (Ester et al., 1996) to extract the density clusters from the 2D projection of the trajectory and identify a representative structure for each cluster. This yields a set of key structures for each trajectory. We obtain a manageable ensemble of structures by eliminating redundant ones using the GROMOS-clustering algorithm, implemented in the GROMACS package (Daura et al., 1999; Hess et al., 2008). To identify the most promising structures in this ensemble, we recycle the AUC score to evaluate the best protein poses according to coevolution. We select the top ten poses for validation of the method.

Evaluating the Representative Structures

To benchmark our method, we compare the predicted alternative structures with those deposited in the PDB. We first check the overlap (Equation 9) between the experimental transition and the sampled one (a value of 1 in the overlap means that the deformation required to move from the reference structure and the simulated alternative conformation is the same than that required to achieve the experimental alternative conformers):

$$\cos\alpha = \frac{|\boldsymbol{\nu}\cdot\mathbf{T}|}{\|\mathbf{T}\|\|\boldsymbol{\nu}\|}, \qquad \text{(Equation 9)}$$

where $\boldsymbol{\nu}$ is the sampled transition vector and $\mathbf{T}$ is the transition vector expected from experimental structures.

The second metric used is the minimum RMSD to target obtained after evaluating the ten proposed alternative conformations.

Supporting Simulations

We run controls to test (1) the impact of the quality of DCA contacts, (2) the significance of our SBM, and (3) the significance of the complete protocol. Regarding (1), we adapt the algorithm by Morcos et al. (2013) into the discrete MD framework. That is, we use pairs from the ranked DCA list directly as single minima at $r_{ij} = 8.0$ Å, complementing a standard Go-like model. To evaluate the importance of the SBM (2), we test the robustness of our protocol by replacing DCA pairs with random pairs. From all possible random pairs, we only consider those at $i - j \geq 5$, and at a distance >12 Å, and with them we reproduce each step of our protocol. To build a multiple-state SBM, we only consider pairs of residues that establish a contact at some point of the pulling trajectory. Finally, to assess the full protocol (3), we compare the ensemble generated with our method with those generated with standard techniques: equilibrium simulations, NMA, and a random-coil model. Random-coil models consist in hard-core interactions (R_{HC} = 2 Å) and bonded interactions to maintain consecutive Cαs at 3.8 Å. NMA-based ensemble is generated by following the top ten normal modes of the source structure individually. We collect 100 structures per eigenvector in both directions, after a relaxation step (Camps et al., 2009; Orellana et al., 2010). Details for equilibrium simulations using SBM can be found elsewhere (Clementi et al., 2000; Emperador et al., 2008a).

p Value Calculation

We run long equilibrium simulations for each case using the Go-like model to describe near-equilibrium protein flexibility. We consider 10,000 structures from the equilibrium trajectory and compute the overlap of each of them to the known transition. Then we use this as background distribution to assess the significance of the overlap obtained using our coevolution-based protocol. We compute the experimental p value as the ratio at which instances with higher overlap values are sampled in the background distribution.

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.str.2015.10.025.

AUTHOR CONTRIBUTIONS

P.S., M.D.F., P.A., and M.O. designed research and wrote the paper. P.S. and M.D.F. developed the code, and ran and analyzed the simulations. A.E. developed the dMD code. R.M. designed the dataset.

ACKNOWLEDGMENTS

P.S. is grateful to Dr. Adam Hospital for help in running large-scale analyses, and Drs. Marcos Villareal and Josep Ll. Gelpi for helpful discussion on FORTRAN implementations of the dMD algorithm. P.S. is a "La Caixa" fellow. M.D.F. is a recipient of the Spanish FPU grant. M.O. is an ICREA Academia Fellow. We acknowledge financial support from the Spanish Ministry of Science (BIO2012-32868 to M.O. and BIO2013-48222 to P.A.), the Catalan Government (SGR_2014 to M.O. and P.A.), and the European Research Council through the grants 291433 (SimDNA) to M.O. and 614944 (SysPharmAD) to P.A.

Received: June 24, 2015
Revised: October 13, 2015
Accepted: October 17, 2015
Published: December 10, 2015

REFERENCES

Bahar, I., Lezon, T.R., Bakan, A., and Shrivastava, I.H. (2010). Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. Chem. Rev. *110*, 1463–1497.

Barducci, A., Bonomi, M., and Parrinello, M. (2011). Metadynamics. Wires Comput. Mol. Sci. *1*, 826–843.

Beckstein, O., Denning, E.J., Perilla, J.R., and Woolf, T.B. (2009). Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open↔closed transitions. J. Mol. Biol. *394*, 160–176.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nuc. Acids Res. *28*, 235–242.

Camps, J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., Cicin-Sain, D., D'Abramo, M., Gelpí, J.L., and Orozco, M. (2009). FlexServ: an integrated tool for the analysis of protein flexibility. Bioinformatics *25*, 1709–1710.

Chen, P.-C., and Hub, J.S. (2014). Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. Biophys. J. *107*, 435–447.

Clementi, C., Nymeyer, H., and Onuchic, J.N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J. Mol. Biol. *298*, 937–953.

Das, A., Gur, M., Cheng, M.H., Jo, S., Bahar, I., and Roux, B. (2014). Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model. PLoS Comput. Biol. *10*, e1003521.

Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W.F., and Mark, A.E. (1999). Peptide folding: when simulation meets experiment. Angew. Chem. Int. Ed. Engl. *38*, 236–240.

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. Nat. Rev. Genet. *14*, 249–261.

Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H., and Shaw, D.E. (2012). Biomolecular simulation: a computational microscope for molecular biology. Annu. Rev. Biophys. *41*, 429–452.

Eisenmesser, E.Z., Bosco, D.A., Akke, M., and Kern, D. (2002). Enzyme dynamics during catalysis. Science *295*, 1520–1523.

Elber, R. (2005). Long-timescale simulation methods. Curr. Opin. Struct. Biol. *15*, 151–156.

Elber, R. (2007). A milestoning study of the kinetics of an allosteric transition: atomically detailed simulations of deoxy scapharca hemoglobin. Biophys. J. *92*, L85–L87.

Elber, R., and West, A. (2010). Atomically detailed simulation of the recovery stroke in myosin by Milestoning. Proc. Natl. Acad. Sci. USA *107*, 5001–5005.

Emperador, A., Carrillo, O., Rueda, M., and Orozco, M. (2008a). Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. Biophys. J. *95*, 2127–2138.

Emperador, A., Meyer, T., and Orozco, M. (2008b). United-atom discrete molecular dynamics of proteins using physics-based potentials. J. Chem. Theor. Comput. *4*, 2001–2010.

Engh, R.A., and Bossemeyer, D. (2002). Structural aspects of protein kinase control—role of conformational flexibility. Pharmacol. Ther. *93*, 99–111.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd *96*, 226–231.

Falke, J.J. (2002). Enzymology. A moving story. Science *295*, 1480–1481.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nuc. Acids Res. *42*, D222–D230.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774–786.

Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. Nature *450*, 964–972.

Henzler-Wildman, K.A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M.A., Petsko, G.A., Karplus, M., et al. (2007). Intrinsic motions along an enzymatic reaction trajectory. Nature *450*, 838–844.

Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theor. Comput. *4*, 435–447.

Hisanaga, Y., Ago, H., Nakagawa, N., Hamada, K., Ida, K., Yamamoto, M., Hori, T., Arii, Y., Sugahara, M., Kuramitsu, S., et al. (2004). Structural basis of the substrate-specific two-step catalysis of long chain fatty acyl-CoA synthetase dimer. J. Biol. Chem. *279*, 31717–31726.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. Cell *149*, 1607–1621.

Hopf, T.A., Morinaga, S., Ihara, S., Touhara, K., Marks, D.S., and Benton, R. (2015). Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. Nat. Commun. *6*, 6077.

Jones, D.T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics *31*, 999–1006.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. USA *110*, 15674–15679.

Khafizov, K., Madrid-Aliste, C., Almo, S.C., and Fiser, A. (2014). Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc. Natl. Acad. Sci. USA *111*, 3733–3738.

Kim, M.K., Jernigan, R.L., and Chirikjian, G.S. (2002). Efficient generation of feasible pathways for protein conformational transitions. Biophys. J. *83*, 1620–1630.

Laio, A., and Parrinello, M. (2002). Escaping free-energy minima. Proc. Natl. Acad. Sci. USA *99*, 12562–12566.

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortíz, A.R. (2005). An analysis of core deformations in protein superfamilies. Biophys. J. *88*, 1291–1299.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

Marks, D.S., Hopf, T.A., and Sander, C. (2012). Protein structure prediction from sequence variation. Nat. Biotechnol. *30*, 1072–1080.

McCammon, J.A., Gelin, B.R., and Karplus, M. (1977). Dynamics of folded proteins. Nature *267*, 585–590.

Michel, M., Hayat, S., Skwark, M.J., Sander, C., Marks, D.S., and Elofsson, A. (2014). PconsFold: improved contact predictions improve protein models. Bioinformatics *30*, i482–i488.

Micheletti, C. (2013). Comparing proteins by their internal dynamics: exploring structure-function relationships beyond static structural alignments. Phys. Life Rev. *10*, 1–26.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. USA *108*, E1293–E1301.

Morcos, F., Jana, B., Hwa, T., and Onuchic, J.N. (2013). Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc. Natl. Acad. Sci. USA *110*, 20533–20538.

Orellana, L., Rueda, M., Ferrer-Costa, C., Lopez-Blanco, J.R., Chacón, P., and Orozco, M. (2010). Approaching elastic network models to molecular dynamics flexibility. J. Chem. Theor. Comput. *6*, 2910–2923.

Orozco, M. (2014). A theoretical view of protein dynamics. Chem. Soc. Rev. *43*, 5051–5066.

Orozco, M., Orellana, L., Hospital, A., Naganathan, A.N., Emperador, A., Carrillo, O., and Gelpí, J.L. (2011). Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings. Adv. Protein Chem. Struct. Biol. *85*, 183–215.

Perilla, J.R., Beckstein, O., Denning, E.J., and Woolf, T.B. (2010). Computing ensembles of transitions from stable states: dynamic importance sampling. J. Comput. Chem. *32*, 196–209.

Proctor, E.A., Ding, F., and Dokholyan, N. (2011). Discrete molecular dynamics. Wires Comput. Mol. Sci. *1*, 80–92.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175.

Seeliger, D., and de Groot, B.L. (2010). Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. PLoS Comput. Biol. *6*, e1000634.

Sfriso, P., Emperador, A., Orellana, L., Hospital, A., Gelpí, J.L., and Orozco, M. (2012). Finding conformational transition pathways from discrete molecular dynamics simulations. J. Chem. Theor. Comput. *8*, 4707–4718.

Sfriso, P., Hospital, A., Emperador, A., and Orozco, M. (2013). Exploration of conformational transition pathways from coarse-grained simulations. Bioinformatics *29*, 1980–1986.

Sfriso, P., Emperador, A., Gelpí, J.L., and Orozco, M. (2015). Discrete molecular dynamics. In Computational MICS: from Quantum to Coarse-Grained Methods Series in Computational Biophysics, M. Fuxreiter, ed. (CRC Press), pp. 339–362.

Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., et al. (2010). Atomic-level characterization of the structural dynamics of proteins. Science *330*, 341–346.

Shimamura, T., Weyand, S., Beckstein, O., Rutherford, N.G., Hadden, J.M., Sharples, D., Sansom, M.S.P., Iwata, S., Henderson, P.J.F., and Cameron, A.D. (2010). Molecular basis of alternating access membrane transport by the sodium-hydantoin transporter Mhp1. Science *328*, 470–473.

Shirvanyants, D., Ding, F., Tsao, D., Ramachandran, S., and Dokholyan, N.V. (2012). Discrete molecular dynamics: an efficient and versatile simulation method for fine protein characterization. J. Phys. Chem. B *116*, 8375–8382.

Stein, A., Rueda, M., Panjkovich, A., Orozco, M., and Aloy, P. (2011). A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. Structure *19*, 881–889.

Taketomi, H., Kan, F., and Go, N. (1988). The effect of amino acid substitution on protein-folding and -unfolding transition studied by computer simulation. Biopolymers *27*, 527–559.

Tozzini, V. (2005). Coarse-grained models for proteins. Curr. Opin. Struct. Biol. *15*, 144–150.

Ueda, Y., Taketomi, H., and Go, N. (1978). Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme. Biopolymers *17*, 1531–1548.

van den Bedem, H., Bhabha, G., Yang, K., Wright, P.E., and Fraser, J.S. (2013). Automated identification of functional dynamic contact networks from X-ray crystallography. Nat. Methods *10*, 896–902.

Velazquez-Muriel, J.A., Rueda, M., Cuesta, I., Pascual-Montano, A., Orozco, M., and Carazo, J.-M. (2009). Comparison of molecular dynamics and superfamily spaces of protein domain deformation. BMC Struct. Biol. *9*, 6.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. USA *106*, 67–72.

Weiss, D.R., and Levitt, M. (2009). Can morphing methods predict intermediate structures? J. Mol. Biol. *385*, 665–674.

Whitford, P.C., Miyashita, O., Levy, Y., and Onuchic, J.N. (2007). Conformational transitions of adenylate kinase: switching by cracking. J. Mol. Biol. *366*, 1661–1671.

Yang, L., Song, G., and Jernigan, R.L. (2007). How well can we understand large-scale protein motions using normal modes of elastic network models? Biophys. J. *93*, 920–929.

Chapter 6: TransAtlas: an integrative database of conformational transitions of proteins

In this work, we present a database of coarse-grained simulations of conformational transitions. We exploited dMD efficiency to simulate almost all conformational transitions in the PDB. After scanning all structures in the Protein Data Bank, we identified structures belonging to the same protein, with at least sequence identity of 80 % and a structural change larger than 2 Å RMSD. We computed trajectories for nearly all-possible combinations of those structures, leading to 750k trajectories. Redundant trajectories are used to extend the sampling on the transition path. For simplicity, trajectories were clustered into 63646 independent transitions. Our physical approach samples plausible conformers without violation of chemical principles. We analysed all transitions paths and obtained the relevant collective variables associated to them, often a requirement for enhanced sampling algorithms. We classified each motion to facilitate the search across the database. Finally, we reconstruct atomistic detail for representative frames to facilitate further atomistic MD calculations. To this end, atomistic models are coupled to our MDWEB server (*251*), where inputs files for major simulating packages are automatically obtained. In summary, we harvested years of methodological development, providing with over an ensemble of conformers distributed over the transition paths for each trajectory, collecting more than 7.5M distinct snapshots.

Title: **TransAtlas: an integrative database of conformational transitions of proteins**

Authors: Pedro Sfriso*, Adam Hospital*, Diana Buitragó, Roberto Mosca, Agustí Emperador, Josep Lluís Gelpí, Patrick Aloy, and Modesto Orozco

Stage: In preparation

Journal:

Type: Research Article

Supplementary Material:

Author Contribution: PS was the main responsible all the work, developed the simulation method and ran the simulations. PS contributed to the writing the paper.

TransAtlas: an integrative database of protein conformational transitions

Pedro Sfriso[1,†], Adam Hospital[1,†], Diana Buitragó[1], Roberto Mosca[1], Agustí Emperador[1], Josep Lluís Gelpí[3], Patrick Aloy[1,2] and Modesto Orozco[1,4*]

[1.] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Barcelona, Spain

[2.] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[3.] Barcelona Supercomputing Center. Barcelona, Spain

[4.] Department of Biochemistry and Molecular Biology, University of Barcelona, Spain

† These authors equally contributed to this work.

* Corresponding author modesto.orozco@irbbarcelona.org

Abstract

We present *TransAtlas*, a comprehensive database of protein conformational transitions. *TransAtlas* contains 64646 independent transitions obtained from the analysis of 750K trajectories covering the nearly entire PDB. Trajectories feeding the database were obtained using a novel coarse-grained method that allows tracing smooth and chemically meaningful transitions between initial and target structures. Trajectories were clustered using a novel approach that makes possible, not only to classify trajectories, but also to define, in an automatic way a set of collective variables, useful for instance, to bias additional simulations. Methods were implemented to generate atomistic structures from all intermediate states in a format ready for atomistic Molecular Dynamics (MD) simulations (equilibrium or biased), including solvated and equilibrated protein structures upon request. The server is freely accessible at http://mmb.irbbarcelona.org/TransAtlas/.

Introduction

Evolution conferred plasticity to proteins as a gateway to broaden their functionality (*1*-*3*). In their sequences, proteins have been coded with dynamical information that define their complex deformation patterns (*4*-*6*). Such flexibility is a common source of noise for structural techniques that traditional experimental setups minimized by design; well-ordered proteins are overrepresented in structural databases. As a consequence, and despite recent experimental advances (*7*, *8*) our knowledge on protein dynamics is limited, and mostly derived from molecular simulations. Atomistic molecular dynamics (MD) is arguably the most popular technique to describe protein flexibility (*9*). The refinement of the algorithms and continuous improvements in the software and hardware are improving the accuracy and the range of applicability of MD simulations (*9*, *10*). However, even with the largest computers MD is still inefficient to trace large conformational transitions (*11*).

MD limitations, as universal sampling technique, are evident when moving from the study of a single protein, to a proteome-scale scenario. The analysis of motions of thousands of proteins is expected to derive principles on protein conformational landscape {Hensen:2012hi}. Thus, while MD-derived databases of the near-equilibrium dynamics of proteins are available (*12*, *13*), generation of equivalent databases for conformational transitions is unfeasible, since it requires to cover, at least 100 times more systems for at least 1000 times longer simulation period. Even assuming that Moore's law stands for the next decades, these calculations will be unfeasible for at least 25 years.

Several approaches have been proposed to overcome current limitations of MD simulations (*11*). A series of methods propose coupling MD sampling engines to biasing scheme that enrich the sampling around the transition pathway (*14*-*17*), reducing the required length of the trajectory. However, these methods require a previous knowledge of the transition pathway, which made them inadequate for proteome-scale simulations. Another family of methods use coarse-grained (CG) approaches (*18*, *19*) where the simplicity of the force-field and the reduction of degrees of freedom yields significant speeds up of simulations. Unfortunately, these methods are not as accurate as atomistic simulations, often being limited to fluctuations around reference structures.

We present here *TransAtlas*, the first proteome-scale database of conformational transitions of proteins. Using a novel CG approach (*20*) powered by a discrete molecular dynamics (dMD) algorithm we explored 750K conformational transitions (63646 of which are unique) deriving 7.5M intermediate snapshots with Cα resolution. From these paths we extracted detailed biophysical information that was used to rationalize and classify the protein motion, as well as, to derive a robust set of collective variables (CVs). CVs capture the conformational change, helping to rationalize protein motion (*21*) and at the same time are very instrumental to bias further simulations. The Cα-resolution transition intermediate were reconstructed to the atomistic detail, generating input files (*22*) for standard MD simulations (either unbiased and biased). The methods and database presented here are publically available at http://mmb.irbbarcelona.org/TransAtlas/.

Results

Data Generation

We scanned the Protein Data Bank (PDB) database (*23*) for proteins showing multiple conformations. We detected pairs of protein chains (> 50 residues) from the same protein and sequence identity >80%. Structural superposition was done for such pairs, selecting those where structure pairs (A and B) displayed root mean square deviations (RMSD) above 2 Å. We retained 17368 structures (PDB chains) generating 447660 conformational transitions. For each conformational transition we ran two trajectories (see Methods) connecting the A→B and B→A states. The collected ensembles were pre-analysed to detect unphysical transitions, where bonds should be broken or created (this might happen in the case of insertion/deletion), and to detect cases where transition is related to trivial tail motions. All these cases were removed from the database, which at the end contains 755440 trajectories.

Trajectory Analysis

Identifying non-redundant transitions

The procedure outlined above generates an ensemble of redundant transitions (for example A→B and A'→B', with A≠B and A'≠B', but A=A' and B=B'), which might be useful to enrich the sampling quality, but that reduces the density of information in the database. Thus, raw transitions were clustered as described in Methods, defining

a total set of 63646 unique transitions. By default *TransAtlas* presents the user a trajectory representative of the cluster, but the user can access and analyse all the rest of trajectories in the cluster.

Capturing key magnitudes of the motion

Trajectories were analysed using a variety of techniques from our FlexServ server (*24*) in order to estimate hinge points, stiffness matrices, variance profiles and residue correlations. As pointed by others (*25*), CG simulations are very useful to derive collective variables (CVs). Internal distances were used to define a set of CVs (see Methods), in a standardized fashion for all cases. Derived CVs are geometrically feasible coordinates that characterize the conformational change (>95% of variance). CVs are suitable to be combined with biasing algorithms, such as Metadynamics (*26*, *27*), steered MD (*28-30*) and Umbrella Sampling (*31*, *32*). Top relevant internal distances can be used in computational pulling experiments, and from there estimate free energy profile and kinetics rates using Jarzynski equality as a base (*21*, *33-35*). CVs can also be useful for energy profiling, ultimately allowing to spot long-time intermediate states in more advanced simulations.

Finding similarity in the transitions

Independent conformational transitions were analysed (see Methods) to detect common transition pathways. We constructed a search tree based on the assigned categories to search for a particular motion type. Quests for similar motions are also possible using one trajectory as reference. *Refinements of the ontology and classification details are still in progress.*

Multiscale modelling

The Cα–resolution intermediate structures visited along the transition path were projected back into the atomistic level using MODELLER (see Methods) to rebuild the side chains. Such models are available for downloading in PDB format, being in total more than 7.5M intermediate structures for all transitions. Models can be used in ligand screening, in search for transient druggable cavities or for protein-protein docking experiments. Liking TransAtlas with our MDWeb server (*22*), we can provide with pre-equilibrated structures in explicit water together with all inputs file for standard MD calculations. Plus, input files for advances simulations like Targeted Molecular Dynamic (*36*) can be also prepared for any two structures from the transition path. *Quality control of structures, and automatic inputs for other advanced simulations techniques is pending.*

Database Architecture

Efficient storage and subsequent retrieval of the large number of conformational transitions computed in this project is achieved using a two-step approach (*12*): 1) a central relational database to store structures and metadata of simulations, as well as analyses results, and 2) a disk-based raw data repository to store generated trajectories. The database design can be divided in information tables (structure and sequence), alignment tables (pair of simulated structures), trajectories tables (containing all metadata from simulations), cluster tables and analyses tables. Information tables of the database are linked with our internal mirrored databases from PDB (*23*), UniProtKb (*37*) and CATH (*38*). Raw trajectories (structures and movies) are stored in the file system following a hierarchical structure, ensuring an efficient data retrieval. The complete file system comprises +15 TB of information.

Data Portal

TransAtlas web server (http://mmb.irbbarcelona.org/TransAtlas) works as a graphical and user-friendly interface to query our conformational transition database. It is built with using new web technologies such as PHP, HTML5, CSS3 and jquery (see help section for details). TransAtlas web portal is divided in two main blocks: **Search** and **Analyses**.

Search

The search section allows finding conformational transitions for a particular protein structure, either from a PDB code, UniProt ID or FASTA sequence. A quick search is always available at the top-right part of the portal to quickly query the database. Computed trajectories can also be browsed using a CATH hierarchy tree, taking as reference the initial structure. Once a particular search has been submitted, a new results tab is populated with all relevant information about the transition. Results are organized according to the transition clusters (see Methods). Representative trajectories display all the information, including a short conformational transition movie. Other transitions in the cluster are listed together with links to simulation metadata and analyses data. Transition trajectories can be downloaded (PDB format) for further local analysis.

Visualization and analyses

The analyses section is enabled after choosing a particular transition. Analyses part is divided in four sections: *summary*, *visualization*, *basic analysis* and *advanced analysis*. *Summary* section shows simulation metadata information such as structures header, compound, resolution and experimental type, simulation time, initial and final RMSD and parameters used in the simulation run. *Visualization* section contains an embedded JSMol applet that offers an interactive view of the conformational transition. The *basic analysis* section shows the complete list of analyses performed for each of the simulations. Description of the different pre-computed analyses offered by the server can be found in the help section. Finally, *advanced analyses* contain a list of links to different platforms to further analyse the generated trajectory. These analysis platforms include: GOdMD (*20*), to generate a new trajectory using dual Go-like dMD simulation; MoDEL (*12*), to find flexibility information extracted from an atomistic molecular dynamics when available; FlexServ (*24*), to automatically run several flexibility analysis for the transition; and MDWeb (*22*), to automatically setup atomistic molecular dynamics simulations for the initial, final or any intermediate structure from the transition path.

Conclusions

In this work, we exploited efficient discrete molecular dynamics simulations to generate thousands of trajectories. Aiming for an integrative approach, coarse-grained transitions path are, in many regards, very informative about the conformational change. Then, where the user decides, fine details coming from atomistic simulation can be added after we provide with the starting structure and standard input files. Sampling can be systematically enhanced adding new departing structures in other conformations. For more exhaustive sampling, or if some parts of the transition path needs refinement, we supplied with input files for advanced simulation methods. Finally, with a broad coverage of conformational transitions captured in the PDB database, we developed tools for a systematically study of protein flexibility.

Acknowledgements

PS and AH are grateful to Mr. J. Alcántara for IT assistance and consulting. PS is a “La Caixa” fellow. MO is an ICREA Academia Fellow. We acknowledge financial

support from the Spanish Ministry of Science (BIO2012-32868 MO and BIO2013-48222 PA), the Catalan Government (SGR_2014 MO and PA), the H2020 program (BioExcel CoE) and Excellerate (MO), and the European Research Council through the grants 291433 (SimDNA) to MO and 614944 (SysPharmAD) to PA.

Author Contributions

PS, AH, PA and MO designed research. PS and AH performed research. PS, RM and PA built the dataset. PS and AE wrote computer programmes. PS and DB implemented the statistical analyses. AH, JLG and MO designed the database. PS, AH, PA and MO wrote the paper.

Methods

Trajectory Generation

We used our GOdMD method to trace the transition path between two known end points (*20*). GOdMD is based on highly efficient discrete Molecular Dynamics (dMD) simulations coupled to an enhanced sampling engine. In the dMD algorithm, particles move at constant velocity until an event (collision) occurs, point at which, velocities are immediately updated. As the advance of the trajectory is based on events instead of integration of Newton's equations of motion (at very short time intervals) simulations are speeded up respect traditional MD (*39-41*). Protein residues are represented with one coarse-grained site centred at Cα position and energy interactions are described with double minima Structure Based Models (SBM;(*42-45*)). Trajectories are smoothly biased towards the target state using a soft-ratchet like algorithm (*46*, *47*). The bias algorithm works as following: a slice of trajectory freely and it accepted with probability p:

$$\text{(Eq. 1)}\quad p(\Delta\varphi) = \begin{cases} 1 & if\ \Delta\varphi \leq 0 \\ e^{-|\gamma\,\Delta\varphi|^2} & if\ \Delta\varphi > 0 \end{cases}$$

where γ is a parameter to control the acceptance rate. $\Delta\varphi$ is an observable that captures the motion of trajectory generated, defined by combination of internal distances. Here, $\Delta\varphi < 0$ means that the proposed slice is moving towards the target structure. Backwards steps, $\Delta\varphi < 0$, can be accepted yielding not necessarily linear trajectories (*48*). If the slice is not accepted, a new one is generated. Repeatedly, the algorithm guarantees net motion towards the target structure, without modifying the

system potential energy landscape. Trajectories were accumulated until the transition path was completed. Afterwards, all trajectories were resized to 100 snapshots to standardize downstream analysis.

Collectivity Index

We adapted the Collectivity Index (κ) proposed by Bruschweiler's group (*49*) to conformational transitions as following:

$$\text{(Eq. 2)} \quad \kappa = \frac{1}{\mathrm{N}} \exp\left(-\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{u}_{\mathrm{i}}^2 \log \mathrm{u}_{\mathrm{i}}^2\right)$$

$$\text{(Eq. 3)} \quad u_i = \frac{1}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N}} \left| \mathrm{r}_{\mathrm{i,B}} - \mathrm{r}_{\mathrm{i,A}} \right|^2}} \left| \mathrm{r}_{\mathrm{i,B}} - \mathrm{r}_{\mathrm{i,A}} \right|^2$$

where N is the number of particles, $\mathrm{r}_{\mathrm{i,A}}$ are the i[th] residue coordinates at the initial structure (A) while $\mathrm{r}_{\mathrm{i,B}}$ are the i[th] residue coordinates at the target structure (B). κ adopts values ranging from 1, when all atoms move exactly the same distance, to 1/N, when only one particle moves.

General Displacement

We constructed a simple magnitude (δ) to evaluate the accumulated displacement over the transition path.

$$\text{(Eq. 4)} \quad \delta = \frac{1}{N\, {R_G}^A} \sum_{i=1}^{N} \sum_{k=1}^{Tf} \left| \mathrm{r}_{\mathrm{i,A}} - \mathrm{r}_{\mathrm{i,k}} \right|$$

N is the number of particles, ${R_G}^A$ is the radius of gyration of the initial conformation (A). Tf is the selected number of snapshots in the trajectory (100), $\mathrm{r}_{\mathrm{i,A}}$ are the coordinates of particle i at the initial conformation while $\mathrm{r}_{\mathrm{i,k}}$ are the coordinates at k[th] snapshot.

Hinge Point Detection

We used the Force Constant approach developed by Lavery's group (*50*) as implemented in our FlexServ server (*24*) to detect hinge points along the trajectories. Force constant for particles was defined with the fluctuations of their positions as following:

$$\text{(Eq. 5)} \quad \theta_i = \frac{Tf}{\sum_t^{Tf} \left(d_i - \overline{d_i}\right)^2}$$

where d_i is the average distance of particle i from other particles j in the protein. $\overline{d_i}$ is the average d_i over the simulation. Again, Tf is the selected number of snapshots in the trajectory (100). Interactions between contiguous Cα are excluded since their distance is constant. Particles within the top 20% of Force Constants distribution were reported as hinge points.

Collective Variables from Partial Least Squares

Collective Variables (CVs) were identified using Partial Least Squares regression (PLS), which detected the fundamental relations between two matrices (X and Y). X, the predictor's matrix, was defined with the values of internal distances over time. Y, the response matrix (vector, in this case), was filled with RMSD to target conformations values over time. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among X values. PLS outputs transformation vectors named components that are in turn, linear combinations of internal distances. After regression, distances were ranked by their weight in each component, with a total number of components selected to preserve >95% of the variance. Selected components are available as high-dimensional reaction coordinates. Plus, for each component, the five distances with larger weight were selected for downstream analysis. In a second step, we removed redundant distances –coming from all selected components– using a correlation cut-off of 0.5. From highly correlated distances, we retained those that 1) participated in the higher ranked component and 2) weight higher within the component. The final number of distances ranges from 1 to 15. In house scripts and package 'pls' in R was used for these computations (*51*).

Cluster Conformational Transitions

We characterized trajectories by two aforementioned magnitudes: the collectivity index and the general displacement (see above). Then, we clustered all trajectories associated to the same protein (sequence identity between initial structures >90%) in the collectivity-displacement space using DBSCAN (*52*). For each cluster, we retrieved the trajectory whose parameters better coincide with the centroid of the cluster. Such trajectory was considered to be a representative one, selected as main entry of the database. We found 63646 representative trajectories that are, within this definition, distinct from each other.

Atomistic Reconstruction

For each trajectory, we selected an ensemble of representative frames by considering one snapshot every 0.5 Å RMSD window respect the initial structure. Reconstruction was carried out with the MODELLER software (*53*, *54*) restraining Cα positions. Latter, 300 Molecular Dynamics steps (in vacuo) were performed to guarantee the suitability of further MD simulations.

Conformational Transition Classification

We classified trajectories in the following categories: oscillations, loop motions, domain motions and complex motions. Tail motions were discarded since are very little informative about protein dynamics. We used three magnitudes to classify conformational transitions: collectivity index, general displacement and the number of PLS components capturing 95% of the variance. (Work in progress)

Molecular Dynamics Inputs Files

We coupled computed transitions to our MDWeb server (*22*), that automatically generates input files for many calculations types. Leading software packages are covered: GROMACS (*55*), NAMD (*56*) and AMBER (*57*) as well as the most popular force fields. For standard MD calculations, users can automatically run up to 0.5 ns, for each structure simulations in our server. For longer simulations, all necessary configuration files are produced and offered in a downloadable file. Also, in an integrative effort, input files for more advanced simulations like Targeted MD or Steered MD can be generated.

References

1. H. Frauenfelder, S. G. Sligar, P. G. Wolynes, The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).
2. K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
3. R. Elber, S. Kirmizialtin, Molecular machines. *Curr. Opin. Struct. Biol.* **23**, 206–211 (2013).
4. C. Micheletti, Comparing proteins by their internal dynamics: Exploring structure–function relationships beyond static structural alignments. *Phys Life Rev* **10**, 1–26 (2013).
5. H. G. Dos Santos, J. Klett, R. Mendez, U. Bastolla, Biochimica et Biophysica Acta. *Biochim. Biophys. Acta* **1834**, 836–846 (2013).
6. P. Sfriso *et al.*, Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins. *Structure*, 1–12 (2015).
7. H. van den Bedem, G. Bhabha, K. Yang, P. E. Wright, J. S. Fraser, Automated identification of functional dynamic contact networks from X-ray crystallography. *Nat. Methods* **10**, 896–902 (2013).
8. F. Schotte *et al.*, Watching a signaling protein function in real time via 100-ps time-resolved Laue crystallography. *Proc. Natl. Acad. Sci. USA* **109**, 19256–19261 (2012).
9. R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, D. E. Shaw, Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* **41**, 429–452 (2012).
10. A. Hospital, J. R. Goñi, M. Orozco, J. L. Gelpí, Molecular dynamics simulations: advances and applications. *AABC* **10**, 37–47 (2015).
11. M. Orozco, A theoretical view of protein dynamics. *Chemical Society Reviews* **43**, 5051–5066 (2014).
12. T. Meyer *et al.*, MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure* **18**, 1399–1409 (2010).
13. M. W. van der Kamp *et al.*, Dynameomics: a comprehensive database of protein dynamics. *Structure* **18**, 423–435 (2010).
14. D. M. Zuckerman, Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.* **40**, 41–62 (2011).
15. R. Elber, Long-timescale simulation methods. *Curr. Opin. Struct. Biol.* **15**, 151–156 (2005).
16. P. Májek, R. Elber, Milestoning without a Reaction Coordinate. *Journal of Chemical Theory and Computation* **6**, 1805–1817 (2010).
17. P. G. Bolhuis, D. Chandler, C. Dellago, P. L. Geissler, Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **53**, 291–318 (2002).
18. H. I. Ingólfsson *et al.*, The power of coarse graining in biomolecular simulations. *WIREs Comput. Mol. Sci.* **4**, 225–248 (2013).
19. F. Ding, N. V. Dokholyan, Simple but predictive protein models. *Trends Biotechnol.* **23**, 450–455 (2005).
20. P. Sfriso, A. Hospital, A. Emperador, M. Orozco, Exploration of conformational transition pathways from coarse-grained simulations. *Bioinformatics* **29**, 1980–1986 (2013).
21. R. B. Best, G. Hummer, Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6732–6737 (2005).
22. A. Hospital *et al.*, MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics* **28**, 1278–1279 (2012).
23. P. W. Rose *et al.*, The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nuc. Acids Res.* **43**, D345–D356 (2015).
24. J. Camps *et al.*, FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics* **25**, 1709–1710 (2009).
25. R. B. Best, G. Hummer, Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. USA* **107**, 1088–1093 (2010).
26. A. Laio, M. Parrinello, Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99**, 12562–12566 (2002).
27. A. Barducci, G. Bussi, M. Parrinello, Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100**, 020603 (2008).
28. S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, K. Schulten, Molecular Dynamics Study of Unbinding of the Avidin-Biotin Complex. *Biophys. J.* **72**, 1568–1581 (1997).
29. B. Isralewitz, M. Gao, K. Schulten, Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.*, 224–230 (2001).
30. H. Grubmüller, B. Heymann, P. Tavan, Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **271**, 997–999 (1996).
31. S. Kumar, J. Rosenberg, D. Bouzida, R. Swendsen, P. Kollman, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **13**, 1011–1021 (1992).
32. G. M. Torrie, J. P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).
33. C. Jarzynski, Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **78**, 2690–

2693 (1997).
34. G. Hummer, A. Szabo, Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proc. Natl. Acad. Sci. USA* **98**, 3658–3661 (2001).
35. G. Hummer, A. Szabo, Kinetics from Nonequilibrium Single-Molecule Pulling Experiments. *Biophys. J.* **85**, 5–15 (2003).
36. J. Schlitter, M. Engels, P. Krüger, Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *Journal of molecular graphics* **12**, 84–89 (1994).
37. The UniProt Consortium, UniProt: a hub for protein information. *Nuc. Acids Res.* **43**, D204–D212 (2015).
38. I. Sillitoe *et al.*, CATH: comprehensive structural and functional annotations for genome sequences. *Nuc. Acids Res.* **43**, D376–D381 (2015).
39. N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. I. Shakhnovich, Discrete molecular dynamics studies of the folding of a protein-like model. *Folding Des.* **3**, 577–587 (1998).
40. E. A. Proctor, F. Ding, N. V. Dokholyan, Discrete molecular dynamics. *WIREs Comput. Mol. Sci.* **1**, 80–92 (2011).
41. P. Sfriso, A. Emperador, J. Gelpí, M. Orozco, in *Series in Computational Biophysics*, (CRC Press, 2014), pp. 339–362.
42. N. Go, T. Noguti, T. Nishikawa, Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* **80**, 3696–3700 (1983).
43. Y. Levy, P. G. Wolynes, J. N. Onuchic, Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. USA* **101**, 511–516 (2004).
44. O. Miyashita, P. G. Wolynes, J. N. Onuchic, Simple Energy Landscape Model for the Kinetics of Functional Transitions in Proteins. *J. Phys. Chem. B* **109**, 1959–1969 (2005).
45. P. C. Whitford, O. Miyashita, Y. Levy, J. N. Onuchic, Molecular Dynamics Studies on the Conformational Transitions of Adenylate Kinase: A Computational Evidence for the Conformational Selection Mechanism. *J. Mol. Biol.* **2013**, 1661–1671 (2007).
46. M. Rueda, E. Cubero, C. A. Laughton, M. Orozco, Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations? *Biophys. J.* **87**, 800–811 (2004).
47. J. R. Perilla, O. Beckstein, E. J. Denning, T. B. Woolf, Computing ensembles of transitions from stable states: Dynamic importance sampling. *Journal of Computational Chemistry* **32**, 196–209 (2010).
48. O. Beckstein, E. J. Denning, J. R. Perilla, T. B. Woolf, Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open↔ closed transitions. *J. Mol. Biol.* **394**, 160–176 (2009).
49. R. Brüschweiler, Collective protein dynamics and nuclear spin relaxation. *J. Chem. Phys.* **102**, 3396 (1995).
50. S. Sacquin-Mora, R. Lavery, Investigating the Local Flexibility of Functional Residues in Hemoproteins. *Biophys. J.* **90**, 2706–2717 (2006).
51. B. Mevik, R. Wehrens, The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal Of Statistical Software* **18**, 1–23 (2007).
52. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise. **96**, 226–231 (1996).
53. N. Eswar *et al.*, in *Current Protocols in Bioinformatics*, (John Wiley & Sons, Inc., 2006).
54. A. Sali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
55. S. Pronk *et al.*, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
56. J. C. Phillips *et al.*, Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781–1802 (2005).
57. D. A. Case *et al.*, AMBER 12. (2012).

Chapter 7: Other Publications

7.1 Dynamics of the large extracellular loop of CD81

Context

Hepatitis C virus recognizes and binds the CD81 membrane protein in its large extracellular loop (LEL) for infection to occur. The virus capsid attaches to the host cell through the E2 receptor and neutral (7.4) pH, but once the virus is invaginated in the vesicle, where there is acidic environment, the binding looses its strength until the virus is freed. We present a mechanism of infection based on the change in pH. Our collaborators crystallized 14 structures of CD81LEL in 3 conformations named open, intermediate and closed. We ran exploratory dMD calculations to connects those states and identify which of the 14 crystallized structures were the most suitable as a starting point for MD calculations. Our Molecular Dynamics study shows that open structures are favoured at acidic pH, supporting the mechanism of action.

Title: The flexibility of the human cellular receptor CD81 large-extracellular-loop serves to enable Hepatitis C virus entry

Authors: Eva S. Cunha, Pedro Sfriso, Adriana L. Rojas, Adam Hospital, Modesto Orozco and Nicola GA Abrescia
Stage: In preparation
Journal:
Type: Research Article
Supplementary Material:

Author Contribution: P.S performed the simulations, analysed the results and contribute to the writing of the paper.

The flexibility of the human cellular receptor CD81 large-extracellular-loop serves to enable Hepatitis C virus entry

Eva S Cunha[1], Pedro Sfriso[2], Adriana L Rojas[1], Adam Hospital[2], Modesto Orozco[2] and Nicola GA Abrescia[1,3]

[1]Structural Biology Unit, CIC bioGUNE, CIBERehd, Derio, Spain

[2]Institute for Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Barcelona, Spain

[3]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Keywords: Hepatitis C virus, CD81 receptor, structural flexibility, virus-receptor interactions

Contact Information: Nicola GA Abrescia, CIC bioGUNE, Vizcaya Technological Park, Bld 800, 48160 Derio (Spain) Tel +34 946572523 Fax +34 946572502

Email nabrescia@cicbiogune.es

List of Abbreviations: HCV, Hepatitis C virus; MD, molecular dynamics; $hCD81_{LEL}$, human CD81 large extracellular loop; Cα, alpha carbon.

Supported by the Spanish *Ministerio de Economia y Competitividad* (BIO2012-32868), by the Catalan SGR, and by the BioExcel and Excellerate EU projects to M.O, by the 'la Caixa' PhD program to P.S. and by the Spanish *Ministerio de Economia y Competitividad* (BFU2012-33947) to N.G.A.A. M.O. is an ICREA Academia Fellow. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under BioStruct-X (grant agreement N° 2460).

ABSTRACT

Hepatitis C virus (HCV) enters into human hepatocytes via interactions with tetraspanin hCD81. Specifically, HCV glycoprotein E2 recognizes the "head" subdomain of the Large-Extracellular-Loop of CD81 ($hCD81_{LEL}$) but the precise mechanism of virus cell attachment and entry remains elusive. The lack of a structure-function model for the CD81 receptor further complicates this understanding.

Here, by combining the structural analysis of a conspicuous number of crystallographically independent $CD81_{LEL}$ molecules (fourteen; ten of which we derived in this study) with molecular dynamics (MD) simulations we show that the dynamism of the $hCD81_{LEL}$ head-subdomain is an inherent property of the receptor. The observed structures of the head-subdomain can be clustered in three conformations, namely *closed*, *intermediate* and *open*: each providing a distinct binding platform for the HCV E2 glycoprotein. Simulations at pH 7.4 and 4 indicate that this conformational variability is modulated by the pH. The crystallised double conformation of the disulfide bridge C157-C175 at the base of the head-subdomain identifies this bond as the molecular zipper of the $hCD81_{LEL}$'s plasticity. Its reduction *in silico* shows a further flexibility at acidic pH. *Conclusion:* The flexibility of the $hCD81_{LEL}$ is inherent to the molecule and it is modulated by the pH and redox conditions enabling virus-receptor interactions to diversely re-engage at acidic endosomal pH. This mechanism rationalizes biochemical data on the priming role of $CD81_{LEL}$ in HCV entry. We propose that this re-engagement favoured by the head-subdomain plasticity renders fusogenic the HCV:$hCD81_{LEL}$ complex. Results presented here will help the structure-based design of virus entry-inhibitors.

INTRODUCTORY STATEMENT

Infection by hepatitis C virus (HCV) affects about 3% of the world population, leading to pathologies such as hepatocarcinoma and liver cirrhosis (1). Currently, there is no clinical vaccine, however direct-acting antivirals (DAAs) with more than 90% success rate are commercially available, at a cost of about $84,000 in the US and £53,000 in the UK per person for a 12-week course of treatment (2).

At the cellular level, HCV enters and starts the infection cycle through the interaction of glycoproteins E1 and E2 with distinct cellular receptors, including tetraspanin CD81 (3, 4). Tetraspanin CD81 is composed of four transmembrane domains, connected by two loops: the Small Extracellular Loop (SEL) and the Large Extracellular Loop (LEL), which is fundamental for interaction with HCV-E2 (5). Importantly, inhibitors abrogating HCV E1/E2 binding to human CD81 have been mapped to the LEL domain ($hCD81_{LEL}$: residues 112-201; Uniprot P60033) and site-directed mutagenesis to $hCD81_{LEL}$ have identified relevant residues for binding (5, 6).

Structural insights into attachment and entry processes into hepatocytes remains are limited despite recent three-dimensional (3D) studies on HCV virus-like and cell-cultured particles (7-9) and on its glycoproteins E2 and E1 (6, 10-12). Differently, the crystal structure of the human $CD81_{LEL}$ – from a decade ago - showed the LEL adopting a five-helix bundle fold composed of a stalk-subdomain (helices α1 and α5) and a head-subdomain (helices α2, α3 and α4) (13, 14)(Fig. 1A). A structural model of the full tetraspanin CD81 molecule has also been generated (15)(Fig. 1B). Across the four $hCD81_{LEL}$ molecules previously crystallised, the head-subdomain shows conformational flexibility within helices α3 and α4 (consensus residues 165-172 and 180-186, respectively), adopting '*closed*' and '*open*' conformations even within the same crystal (13, 14). This variability has casted some doubt as to whether these conformations are physiologically relevant (16-19). By NMR, $hCD81_{LEL}$ helix α4 appears to be disordered in solution (at pH 7.0), whereas previous energy minimization studies suggested that helix α4 is ordered and forming a narrow cleft with helix α3 (*closed* conformation) (18, 19).

Biochemical studies on the interaction between $hCD81_{LEL}$ and HCV E2 and/or HCV viral particles indicate that the $hCD81_{LEL}$ primes the HCV glycoproteins for low pH-dependent fusion, however, this mechanism of action remains unclear (5, 6, 11, 12, 19-22).

Since HCV entry is a promising alternative route for therapy, we investigated the $hCD81_{LEL}$ head-subdomain structural flexibility by X-ray crystallography and by molecular dynamics (MD) simulations.

We crystallized the $hCD81_{LEL}$ molecule in four novel crystal forms, which captured a total of ten $hCD81_{LEL}$ molecular conformations. Structural analysis of all $hCD81_{LEL}$

molecules shows the head-subdomain as the most flexible region of the protein. Moreover, the disulfide bridge formed by C157-C175 upon which the head-subdomain module hinges, displayed multiple conformations. Careful analysis of the geometry of helices α3 and α4 across the fourteen crystallised $hCD81_{LEL}$ molecules distinguishes a conformation *intermediate* between the *closed* and *open* ones. Accordingly, MD simulations unequivocally illustrate that $hCD81_{LEL}$ is a highly dynamical molecule. *In silico* pH titration studies support a pH dependency of the head-subdomain conformations, with the open conformation favoured in acidic conditions. The reduction *in silico* of the disulfide bridge C157-C175 produces an increased flexibility and opening of the head-subdomain identifying this bond as the molecular zipper of the head module.

We propose that the observed $hCD81_{LEL}$ conformational flexibility is exploited by HCV at cell entry with the environmental pH and redox changes modulating the conformational space visited by the head-subdomain thus allowing the virus-receptor complex to transit into the fusogenic state. Our study provides significant insights into the influential role of $hCD81_{LEL}$ on the HCV attachment mechanism and it aids the rational design of new antivirals blocking HCV entry.

MATERIAL AND METHODS

Cloning and protein purification

The $hCD81_{LEL}$ gene was cloned into the pHLSec vector (this introduced the three extra residues ETG at the N-terminus and $GTKH_6$ at the C-terminus) and the protein was transiently expressed as a secreted, highly soluble protein in HEK293T mammalian cells and purified as previously described (23). Briefly, after Ni-affinity chromatography, the protein was concentrated, buffer-exchanged into 20mM Tris pH 7.2 and 150 mM NaCl and loaded onto a Superdex 75/HR 10/30 column (GE-Healthcare); the eluted fractions corresponding to the $hCD81_{LEL}$ dimer were concentrated to 10 mg/ml using 10 kDa MW-cutoff Vivaspin concentrator.

Crystallization, data collection and processing

More than 1000 commercial crystallization conditions were screened for $CD81_{LEL}$

crystallization using a Mosquito Crystal robot (TTP LabTech). Vapour diffusion sitting drops were set-up with volumes between 280 and 200 nl at protein:precipitant ratios of 2:1 and 1:1. The grown crystals were cryo-protected in either 20% glycerol or ethylene glycol and flash-frozen in liquid nitrogen prior X-ray diffraction data collection at European Synchrotron Radiation Facility (France) and at Diamond Light Source (UK) synchrotrons. Processing of diffraction images was performed either with the HKL2000 suite (24) or with XDS (25). Then each dataset was scaled and merged in CCP4-Aimless (26); four novel crystal forms were obtained ($P6_222$, $P3_1$, $C2$ and $C222_1$) in addition to the previously found ($P2_1$ and $P6_4$) (Table 1).

Some crystals in $P2_1$ were grown also in presence of synthetic claudin I long-extracellular-loop (CLDN1-EL1, residues 29-53) tagged with fluorescein (CASLO, Aarhus, Denmark) and although crystals appeared yellow, no ordered CLDN1-EL1 was seen in the corresponding electron density. Also, benzyl salicylate and fexofenadine were used in co-crystallization experiments (*e.g.* in the $P6_4$ and $C222_1$ crystal forms) but neither ligand was found in the structures.

Crystal structure determination and refinement

All structures were solved by molecular replacement technique in Phaser software in CCP4 (26) using PDB ID 1G8Q chain A as a search model (13). Alternate rounds of refinement in Phenix software (27) and manual model rebuilding in COOT (28) led to the final deposited models (Table 1). Overall starting B-factors were assigned by the Wilson-plot analysis. All structures were refined with individual isotropic B factors and TLS with the exception of the crystal form $P2_1$ (1.3 Å) where the B-factor were refined anisotropically. For the $C2$ crystal form, the NCS were maintained at all stages of the refinement except for the regions from residues 137 to 142 and 161 to 187. In the case of the $C222_1$ crystal form refinement was carried out using at all times NCS and the Deformable Elastic Network (DEN) as implemented in Phenix (27). Final statistics for the refined models is shown in Table 1.

Structural analysis

The secondary structure of the individual X-ray structures was assigned in STRIDE (29) and the geometry of the head-subdomain containing helix α3 and α4 was assessed using the inter-helical angle (θ) adapting the Kahns's method in Qhelix (30) and the radius of gyration R_g which is a robust indicator of the overall shape of the head-module. First, we obtained the line that best fitted to each helix axis (using the consensus residues 165-172 and 180-186) and then computed the angle between them:

$$\cos\theta = \frac{|\boldsymbol{I_3} \cdot \boldsymbol{I_4}|}{\|\boldsymbol{I_3}\|\|\boldsymbol{I_4}\|}$$

where I_3, I_4 are respectively the helical axes of α_3 and α_4 whereas the R_g is defined as:

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\vec{r}_i - \vec{r}_{CM})^2}$$

where $\vec{r}_i$ are the spatial coordinates of each particle, $\vec{r}_{CM}$ is the centre of mass and n runs for Cα residues 165-172 and 180-186 giving total of N particles.

The above two metrics were also used to analyse the simulations and MD trajectories (see below)

The packing analysis of each crystallised molecule was determined by calculating the accessible surface area difference (Δ-ASA) in the context of crystal packing *versus* the molecule on its own using the AREAIMOL software in CCP4 (26). The pH-dependent properties of these molecules were examined using PROPKA software and GUI version (31) and the electrostatic isopotential surfaces (at pH 7.4 and pH 4.0) were calculated using PDB2PQR and APBS software (http://nbcr-222.ucsd.edu/pdb2pqr_2.0.0/) and visualised in Pymol (https://www.pymol.org/).

Exploratory discrete molecular dynamics simulations

The GOdMD software (32) was used to trace the conformational transitions between all combination of *open* conformations (molecules *12* and *13*) and *closed* conformations (molecules *1-7*, *9*, *10* and *14*) (Fig. 1A). This led to 40 trajectories (*open* → *closed* and *closed* → *open*) that generated an ensemble of more than 50000 structures. The

conformational space spanned by these simulations was analysed using Principal Component Analysis (PCA). We selected the most representative structures (molecules *1*, *7*, *11*, *12* and *13*) as the starting point for follow up MD simulations (see Fig. S1).

Molecular dynamics of CD81$_{LEL}$

Proteins were titrated, neutralized, hydrated, minimized, heated and equilibrated using standard protocols (33, 34). Simulations were performed using AMBER 14, parm99SB force field with ILDN side-chain torsion corrections and explicit solvent (TIP3P model) (35, 36). After structures setup (34), an equilibration step was applied to the resulting systems, which were allowed to relax for 12 ns. These equilibrated structures were then used as starting points for 1 μ s production trajectories. Trajectories were performed at constant pressure (1 atm) and temperature (300 K) using standard coupling schemes (33, 34). For each of the proteins, simulations were run in two pH conditions: 7.4 and 4.0. Proper protonation state of ionizable residues was set using standard protocols and in-house software for those residues requiring more accurate inspection (37, 38). All trajectories were ran in duplicate, using as input structures different snapshot collected from the equilibration after assigning a different set of random velocities. Hence, 20 simulations were performed (5 structures x 2 pH conditions x 2 replicas) with a total accumulated time of 20 μs of hCD81$_{LEL}$ dynamics.

To prove the influence of the C157-C175 disulphide bond on the head-subdomain plasticity, we ran MD simulations where the S-S bridge was removed. To this purpose, we used three different starting conformations (mol-*1*, mol-*11*, and mol-*13*; Figs. 1A) and two pH conditions: 7.4 and 4.0.

RESULTS

hCD81$_{LEL}$ crystallizes in four new crystal forms

The four new crystal forms of hCD81$_{LEL}$ belonged to space groups $P6_222$, $P3_1$, $C2$, and $C222_1$, diffracting to 2.4, 2.0, 3.1 and 5.0 Å resolution, respectively. All crystals grew at T=21 ^{o}C and pH $\leq$ 6 except those in $C222_1$, grown at pH 7.5 (Table 1). We also

reproduced the crystals originally found in $P2_1$ and $P6_4$ crystal forms [PDB IDs 1G8Q and 1IV5 (13, 14)] improving their resolution to 1.3 Å and 2.0 Å, respectively (Fig. 1A and Table 1). Through structure solution of the $P6_222$, $P3_1$ and $C2$ crystal forms ten new crystallographically independent conformations of $hCD81_{LEL}$ (numbered *1-10*) were added to the four molecules in the $P2_1$ and $P6_4$ structures previously described and referred to with the numbers *11-12* and *13-14* (Table 1 and Fig. 2A). The low-resolution $C222_1$ crystal form (~5 Å) had four $hCD81_{LEL}$ molecules in the asymmetric unit but due to the low resolution these molecules were not included in the structural analysis. In all crystals $hCD81_{LEL}$ molecules form dimers (Fig. 1A) as in solution (data not shown).

The $hCD81_{LEL}$ head-subdomain structures cluster into three conformational classes

Pairwise superimposition across the fourteen $hCD81_{LEL}$ conformations with SHP (39) showed an average root-mean-square-deviation, <rmsd>, of 1.12 Å and σ_{rmsd}= 0.49 Å over ~85 Cαs (Fig. 2B). We then split each monomer into stalk- and head- subdomains. We found the stalk subdomain to be on average rigid, with <rmsd>=0.59 Å and σ_{rmsd}= 0.18 Å over ~52 Cαs whereas the head-subdomain to be most dynamic module within the $hCD81_{LEL}$ with an <rmsd> of 1.44 Å and σ_{rmsd}= 0.89 Å over just ~30 Cαs (Fig. 2B and Movie S1).

To structurally analyse the different head-subdomain conformers we first assigned the secondary structure to the fourteen $hCD81_{LEL}$ atomic models (Fig. 2A). Then the inter-helical angle (θ) between helices α3 and α4 and the corresponding radius of gyration (R_g) were estimated. The secondary structure algorithm mainly assigns a α-helix to residues 161-170 (helix α3) whereas residues 181-186 (helix α4) differ from a strict α-helix as in the case of mol-*1* ($P6_222$) to a mixture of turn, bend, 3_{10}-helix and α-helix across other $CD81_{LEL}$ molecules (Fig. 2A). Specifically, two out of the five molecules composing the asymmetric unit of the $C2$ crystal form loose helicity from residue 181 whereas helix α3 in mol-*9* is the shortest (Fig. 2A).

The measured inter-helical angles between α3 and α4 ranges from 41^o to 179^o indicating a different relative arrangement of the two helices (with values close to 180^o for helices with almost parallel axes). This study coupled with the analysis of the R_g allows

visualizing the conformational spread of the fourteen structures (Fig. 2B-C). Based on the above geometrical parameters the head-subdomain *closed* conformation can be defined by R_g < 6.7 Å, and a large inter-helical angle, $150^o < \theta < 180^o$, whereas the *open* conformation by R_g > 7.5 Å. However, specific combinations of R_g and θ further modulate the widening of the α3 and α4 groove. This modulation defines the *intermediate* conformation seen with $\theta = 41^o$ and a $R_g \cong 7.7$ Å or the *open* ones with $\theta = 45^o$ and a R_g = 9.5 Å or $\theta = 134^o$ and a $R_g \cong 7.6$ Å (Fig. 2C, below). In this latter case the α3 and α4 helices are almost parallel but their inter-helical groove is wider than in the *closed* structure. The majority of the crystallised molecules adopt a *closed* conformation (Fig. 2C).

Furthermore, the hinge-angle of the head-subdomain over the stalk-subdomain varies across molecules, increasing further the plasticity of the $hCD81_{LEL}$ ectodomain. For example, taking a *closed* molecule (mol-*1*) as a reference and using the stalk as pivot, the head-subdomain hinge of molecule 6 differs by ~2^o, while difference in the same hinge angle for mol-*11* is ~16^o (Fig. 2A-C).

Crystal packing marginally influences the head-subdomain conformations

Comparison of the molecular packing density through the estimation of the water content (Vs) across the crystal forms (a higher solvent content implies more loosely packed molecules) excludes a correlation between the head-subdomain flexibility and the amount of solvent present in each crystal. The head-subdomains in molecules in *C*2 (Vs= ~69%) and those in $P6_222$ and $P3_1$ (Vs= ~42%) possess mainly a *closed* conformation whereas molecules within the $P2_1$ crystal, with the lowest Vs=35.6%, and $P6_4$ (Vs=48.4%) harbour both *closed*/*open* and *closed*/*intermediate* head-subdomain conformations, respectively. The four $CD81_{LEL}$ molecules packing into the crystal form $C222_1$ with the largest Vs=74.6% and diffracting only to 5 Å appear all to display a *closed* conformation (data not shown).

Also to rule out the possibility of crystal contacts being the major responsible for the conformational variability observed in the head-subdomain, we analysed, per residue, the variation of the surface area accessible (Δ-ASA) to solvent of each of the head-

subdomains when alone and when in the context of neighbouring molecules. A large variation would imply a marked environmental difference. With the caveat that flexible regions can always be affected by the surroundings (40), the Δ-ASA analysis and relative comparison on the α3 and α4 helices indicates that a defined head-subdomain conformation is independent from the amount of contacts that overall the domain establishes with neighbours (Fig. 3). This is even more evident across the $CD81_{LEL}$ molecules in *closed* conformation whose variation of the surface area accessible ranges from about -198 $Å^2$ to about -2800 $Å^2$ (red-dots in Fig. 3). Moreover specific residue differences are present in each head-subdomain conformation; in one case only, the visual inspection together with the STRIDE assignment (Fig. 2A) and the above analysis pinpoint at residues 165 and 166 in α3 (mol-*9*) as to closely engaging with symmetry-related residue 180 in α4 and thus possibly influencing the α3 helicity propensity (Fig. 3). Overall this analysis challenges the idea that these conformations (specifically those referring to α4) are consequence of packing artefacts since each molecule sees different environments and suggests that the head-subdomain structure is inherently flexible and thus structurally responsive to the environment.

Disulfide bonds: molecular zippers of the hCD81$_{LEL}$ head-subdomain

The structural plasticity of proteins from X-ray data manifested through alternate side-chain and/or backbone conformations is detectable at resolution better than 2.0 Å. The 1.3Å resolution obtained for the crystal in space group $P2_1$ distinguishes two conformations of the C157-C175 disulfide bond critically located at the base of the head-subdomain. The clear visualization of this heterogeneity underpins its pivotal structural role (Fig. 2C-D).

The cysteine rotamers for S-S bridge in conformation-I adopt the most favourable orientation for C157 with 50% of occurrence in the rotamer database and a side chain angle chi 1 ($\chi1$) of -65^o and the second most favourable one for C175 (26% frequency and $\chi1$= -177^o) (Fig. 2D, top). In the alternative S-S bridge conformation-II (Fig. 2D, bottom), C157 rotamer adopts the second most probable structure whereas C175 nearly the most favourable one. The overall relative occupancies of conformation I *versus* II is

~55% and ~45%, respectively. Interestingly, in the other hCD81$_{LEL}$ molecule (mol-*11*) with the head-subdomain in the *intermediate* conformation only C175 shows a double conformation as almost flickering between two close rotamer structures with relative occupancy of 0.29 and 0.71 and both leading to a S-S bridge to the more static C157 (Fig. 2D, bottom). None of the remaining crystal forms achieved a resolution higher enough for which alternate conformations could be refined, and all displayed the S-S conformation-I. Moreover, comparison of the B factors across the residues within the head-subdomain of all structures clearly show the 'zipping' role played by the S-S bridge with relative lower B factor values for the C175 and C157 (Fig. S2).

Interestingly, beneath this disulfide bond and at ~3.5 Å distance from the C157 Cα lays the aromatic ring of H151 within α1 helix. In all fourteen structures H151 is hydrogen-bonded with Y127 in helix α2. At 1.3 Å resolution it is not possible to determine experimentally the protonation state of H151 but assignment *in silico* of protonation states identifies residues H151 and Y127 as those with the most shifted values from the model pKa values (-2.1 and 3.9 pH units, respectively) (Fig. 2E). Noticeably, this analysis also suggests that H151 provides the highest stabilizing contribution (dG 2.7) to the free energy needed to unfold the molecule at given pH 7.4 whereas at pH 4.0 its stabilizing contribution is practically nil (dG 0.2) (see below).

Each head-subdomain conformation displays distinct binding platforms

The structural plasticity of the head-subdomain implies distinct landscapes for Hepatitis C virus-receptor binding which can be further influenced by environmental conditions. The electrostatic isopotential surface calculated at pH 7.4 differs across the *closed*, *intermediate* and *open* head-subdomain conformations (Fig. 2F, top).

At pH 7.4 the *closed* conformation displays a hydrophobic surface contributed by exposed residues L165, V169, L170 on helix α3 and V181 and I185 on α4. The opening of the groove between helices α3 and α4 exposes buried charged residues such as N173 and N184. The structural variations on the α4 helix and changes in the helices groove modulate the specific electrostatic properties of each head-subdomain conformer. Moreover when the electrostatic isopotential surface is calculated at pH 4.0 further modulation towards a more positive charge distribution is noticeable (Fig. 2F, bottom).

These observations derived from the $CD81_{LEL}$ crystal structures illustrate how the inherent head-subdomain flexibility combined with environmental conditions provides different binding molecular properties and engagement modes to the HCV-E2 glycoprotein.

Microsecond-timescale molecular dynamics reproduce the observed $hCD81_{LEL}$ conformational states

MD simulations have proven an essential tool to structural biology (41, 42). Also, MD combined with experimental data added atomistic functional pathways and probed the topological space of proteins (43).

Thus, we conducted a MD study to explore the dynamical properties of $hCD81_{LEL}$ considered in its monomeric state as from previous studies (6, 21, 22). For the selection of the best initial models out the fourteen solved X-ray structures for atomistic MD study, an exploratory coarse-grained simulation was performed (32). This protocol identified molecules *1*, *7*, *11*, *12* and *13* as representative class structures (Fig. 2A-B). Then using state-of-the-art MD protocols over 20 microseconds of $hCD81_{LEL}$ dynamics were accumulated. This is to our knowledge the longest simulation of $hCD81_{LEL}$ published to date. Simulations unequivocally show that the most mobile part of the $CD81_{LEL}$ is the head-subdomain, while the stalk-subdomain is stable along the trajectories (Fig 4A). Notably, the dynamic properties of the system are mostly defined by helices $\alpha3$, $\alpha4$ responsible for HCV E2: CD81 interactions, and by the connecting loop between $\alpha1$ and $\alpha2$ helices (aa 132-136) (Fig. 4B). Furthermore, this study illustrates that the $CD81_{LEL}$ molecular conformations framed within the crystal structures are not oddities caused by crystal contacts but meaningful conformations in solution as consistently visited during the exploration of the conformational space (Fig. 4C).

In silico protonation of $hCD81_{LEL}$ correlates pH and head-subdomain conformations

The lack of correlation between crystal packing forces and observed $CD81_{LEL}$ structures and the changes in electrostatics across the framed head-subdomain conformations at

different pHs (Fig. 2E) raised the question whether micro-environmental conditions could modulate the $CD81_{LEL}$ plasticity.

Using MD simulations we investigated the possible dependency of the head-subdomain flexibility from pH conditions at 7.4 and 4.0, mimicking respectively the cellular and endosomal pH, by *in-silico* pH titration (44). Analysis of the evolution of the molecular trajectories at neutral and acidic pH was carried out using the R_g of $\alpha 3$ and $\alpha 4$ helixes as monitoring parameter for the closing or opening of the head-subdomain.

We found that at acidic pH $hCD81_{LEL}$ shifts its population towards *open* conformations (as the distribution extends to higher R_g values; regardless of the starting point structure (Fig 4D). Furthermore, at pH 4.0, protonation of H151, underneath the C157-C175 bond, amplifies the displayed motility and disrupts the crystallographically observed H151-Y127 hydrogen-bond supporting this region as the core sensing switch (Fig. 2D). At neutral pH the population of $hCD81_{LEL}$ favours *closed* conformers (Fig. 4D and Movie S2).

To explore also the impact of redox conditions in the conformational state of the protein, we reduced *in silico* the disulfide bridge in each of the head-subdomain conformer and repeated the MD at pH 7.4 and pH 4.0. At pH 7.4 the presence or absence of the C157-C175 bridge has little effect on the head sub-domain whereas at pH 4.0 the head-subdomain module become unstable in all three conformers as monitored by the increase distance between the two sulphur atoms (Fig. 4E).

These results strongly imply that pH and redox environmental conditions are key factors influencing the conformational polymorphism of the $CD81_{LEL}$ region involved in the HCV binding.

DISCUSSION

The $hCD81_{LEL}$ head-subdomain is a molecular 'gymnast'

Accurate structural data are essential for structure-based drug design, and crystal structures are currently a reliable source for determining active site geometry, modes of protein-protein interaction, protein-ligand interaction, etc (45). Furthermore, alternative physiological active conformations can also be trapped and studied within crystals (46).

Our data provide insights into the conformational dynamism of the $CD81_{LEL}$ that together with *in silico* modelling allow us to better understand its implications in HCV cell recognition, attachment and entry which can be instrumental for the development of novel therapeutic strategies (47).

In all analysed crystal structures the residues corresponding to helix α4 within the head-subdomain and responsible for binding to HCV E2 glycoprotein, were clearly traced in the electron density although their secondary structure varies across molecules (Fig. 2A). The analysis of contacts across fourteen $CD81_{LEL}$ molecules of five different crystal forms, and state-of-the-art computer simulations demonstrate that the dynamism of head-subdomain is an inherent property of the $hCD81_{LEL}$ providing the atomic detail of this variability (Figs. 3 and 4). These crystal structures frame different dynamic states of the head-subdomain, which become invisible by solution NMR studies (19).

Our analysis supports $CD81_{LEL}$'s structure as composed at least by two dynamic parts: (*i*) the structural variability of the head-subdomain and (*ii*) the head-subdomain module hinging over the stalk domain. Previous studies and inspection of the full-length CD81 predicted model favour a third dynamic component, the ectodomain module, possibly pivoting over the transmembrane (TM) helices via glycine residues located almost at matching height (*eg.* G30 on TM1, G61 on TM2, G109 and G112 on TM3 and G200 and G206 on TM4) (15, 19). This modularity confers to CD81 a higher conformational plasticity than a solely disorder-order transition in the head-subdomain (Fig. 5A).

Thus, we propose that the dependency of the head-subdomain polymorphism from the pH and redox conditions has important implications in CD81 tetraspanin cellular functions beyond its role as HCV receptor (see below).

The $hCD81_{LEL}$ head-subdomain structural plasticity serves to HCV entry

Upon attachment of HCV to CD81, the complex translocates to the tight junction where CD81 then binds to Claudin-I for successful endocytotic uptake (48). Internalization may be aided by additional co-factor proteins (49). However, the molecular mechanisms governing these transition states of HCV cell entry are far from being understood.

Unlike other *Flaviviridae* members that have been shown to possess class II fusion glycoproteins fully capable of membrane fusion at low-pH, HCV glycoprotein E2 is

unusual in this respect requiring CD81 and recent evidence also point to HCV-E1 glycoprotein involvement into the fusion process (12, 20, 50). Antibody studies suggest that tightly coupling of the back layer domain of E2 with CD81 at binding implies that structural rearrangements must occur for transition from a pre-fusion state to the fusogenic one (6, 21). Also previous elegant biochemical studies have shown that the binding of $CD81_{LEL}$ to E2 is a necessary-but-not-sufficient condition for internalization and further endosomal fusion, suggesting that the environmental acidification primes the initially formed E2:$CD81_{LEL}$ complex to become fusogenic (20).

Our combined structural and dynamic studies on $CD81_{LEL}$ head-subdomain frame this cellular ectodomain conformation responsive to pH changes and redox conditions, with the *closed* conformation favoured at pH 7.4 and the *open* one at pH 4.0. The observed intrinsic polymorphism of the C157-C175 disulfide bridge, the conserved hydrogen bonding between H151-Y127 and the MD protonation studies also identify these residues as candidate molecular sensors of environmental changes.

We envisage that these molecular properties intrinsic to tetraspanin CD81 cellular functions are hijacked and exploited by HCV at entry. In this view, the population of the $hCD81_{LEL}$ head-subdomain conformers at the hepatocyte surface would be biased towards the *closed* structure since the extra-cellular milieu is a basic (pH 7.4) and oxidizing environment. Therefore HCV would initially bind to the $CD81_{LEL}$ *closed* conformation. However, upon endocytotic uptake with the acidic and reductive conditions of the endosome, the breaking of the above S-S bridge and hydrogen bond would free the head-subdomain to structurally re-arrange with the HCV-E2 leading to the fusogenic virus-receptor complex (Fig. 5B). This molecular mechanism rationalizes the priming role formerly ascribed to the $CD81_{LEL}$ domain (20).

In conclusion, our study shows that the $hCD81_{LEL}$ head-subdomain plasticity is inherent to the CD81 molecule and exploitable by HCV at entry and it provides the molecular understanding behind the fusion-dependent rearrangements of the virus-$CD81_{LEL}$ complex previously postulated. In this view the design of entry-inhibitors will benefit from considering the conformational variability of the $hCD81_{LEL}$ head-subdomain.

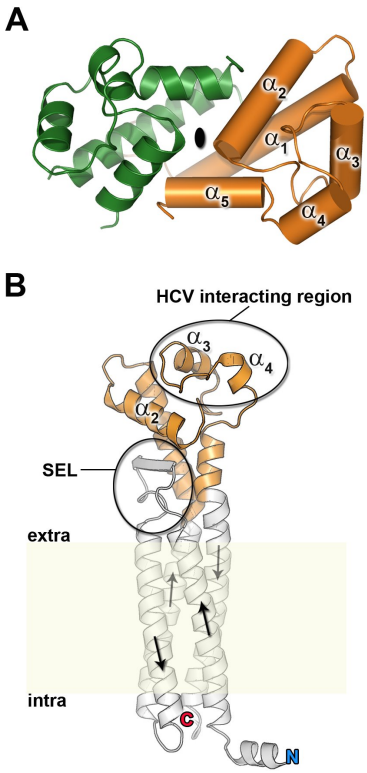
ACKNOWLEDGEMENTS

We are grateful to Radu Aricescu (Oxford University) for supplying the pHLsec vector and for useful suggestions on mammalian protein expression, to Magdalena Wojtas (CIC bioGUNE) for assisting in cloning $CD81_{LEL}$ and, to Marina Ondiviela and Diego Charro (CIC bioGUNE) for protein production. Marina Ondiviela and Pietro Roversi (Ikerbasque Visiting Fellow) are thanked for providing insightful help in protein purification, crystallization and data collection. We also thank Hani Boshra (CIC bioGUNE) for discussion and critical reading of the manuscript.We acknowledge the Diamond Light Source and the European Synchrotron Radiation Facility (ESRF) for provision of synchrotron facilities. We also thank the staff at the beamlines I02/I03 at Diamond-UK and ID-29/ID23-1 at ESRF-France.

Accession codes: Coordinates and structures factors have been deposited in the Protein Data Bank with the following codes AAAA ($P6_222$ form), BBBB ($P3_1$ form), CCCC ($C2$ form), DDDD ($C222_1$ form), EEEE ($P2_1$ form) and FFFF ($P6_4$ form).

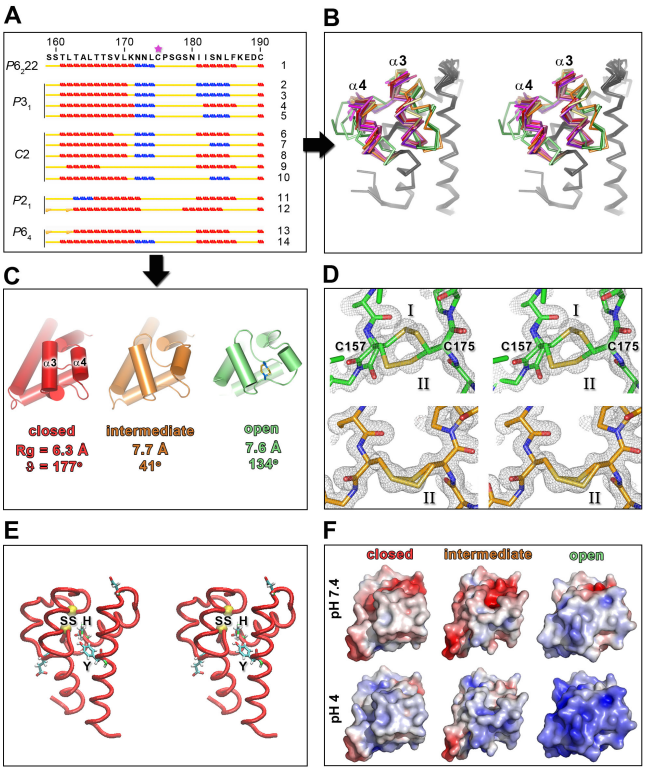
FIGURES

Figure 1. Human CD81 molecule.



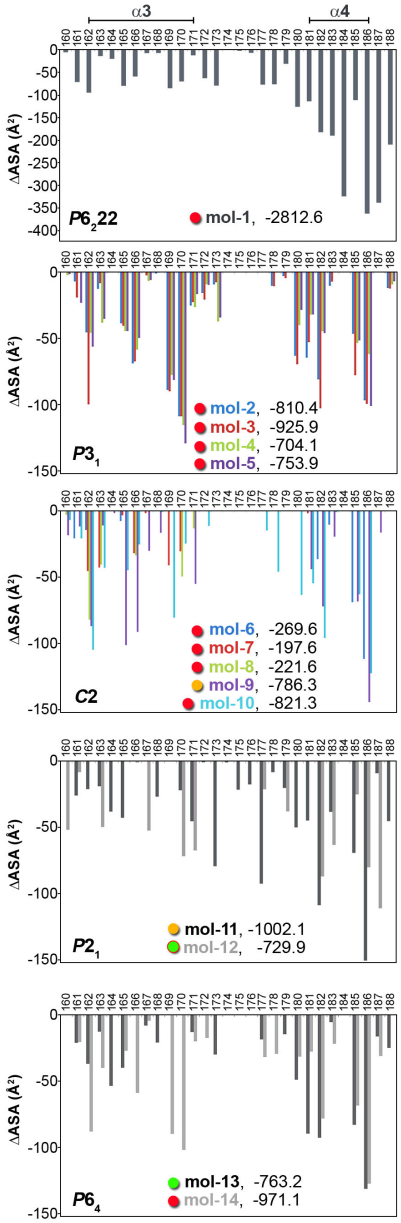
(A) Overall view of the dimeric arrangement of $hCD81_{LEL}$ molecules in the current high-resolution $P2_1$ crystal and present as common dimeric motif (oval symbol, two-fold symmetry axis) across all six crystal forms (Table 1); one molecule as green cartoon and the other as orange cylinders schematically showing the five-helix bundle topology (from STRIDE secondary structure assignment: α1=116-136/116-136; α2:141-154/142-154, α3=166-172/163-172; α4=181-186/179-184; α5=190-199/190-199).
(B) Chimeric montage of the above orange $hCD81_{LEL}$ crystal structure and the predicted full tetraspanin CD81 model in white (PDB ID 2AVZ (15)); represented as cartoon with black-arrows marking the direction of the transmembrane helices from the N-terminus (N, blue) to the C-terminus (C, red), with black circles the short-extra-cellular-loop (SEL) and helices α3 and α4 responsible for Hepatitis C virus binding, respectively; in pale-cream the schematic membrane separating the intra- and extra-cellular regions.

Figure 2. Conformational variability of the hCD81$_{LEL}$ head-subdomain.



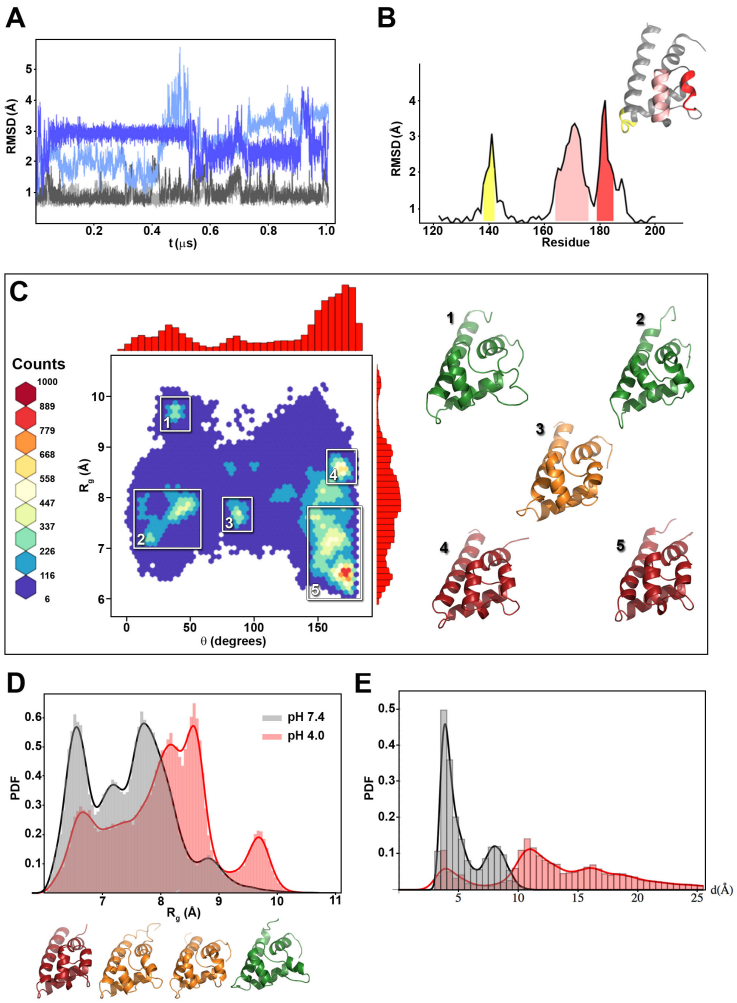
(A) Graphic representation of the secondary structure assignment of the fourteen X-ray structures; red helices α-helix, blue helices 3_{10} helix, yellow bar turn or coil.
(B) Stereo-view of the superposition of the fourteen hCD81$_{LEL}$ molecules showing the head-subdomain dynamism; the stalk-subdomain and α2 helix (black) with residues from 156 to 186 containing the helices α3 and α4 coloured differently for each molecule (see also Movie S1). (C) Cylinder representation of hCD81$_{LEL}$ class representatives based on the radius-of-gyration (R_g) inter-helical angle (θ) between α3 and α4 helices and colour coded as in (A) with red, orange and green representing the *closed*, *intermediate* and *open* conformations, respectively. In the open conformer the two observed conformations of the C157-C175 disulfide bond (yellow) are depicted in stick. (D) View of the C157-C175 disulfide bridge [rotated 90^o clockwise from panel (C)] with top inset showing as stereo-view the 2Fo-Fc electron density (contoured at 1.0 σ level white mesh) corresponding to the conformations I and II of C157-C175 in the open head-subdomain (mol-*12*); below as above but the alternate conformation for C175 only in the intermediate head-subdomain (mol-*11*); (E) Stereo-view of the location of the four ionizable residues (in stick representation and colour-coded according to atom) with the two most shifted pKa values in hCD81$_{LEL}$ being hydrogen-bond interacting H151 and Y127 residues (at the core of the molecule) followed by the more exposed residues D138 and E188 (respectively -1.7 and 1.3 pH units) ; the yellow sphere identify C157 and C175 forming the S-S bridge.
(F) Electrostatic isopotential surfaces of the three major head-subdomain conformations (as C) countered at levels of -5 kT/e (red) and 5 kT/e (blue) calculated at pH 7.4 (top) and at the pH 4.0 (bottom).

Figure 3. Analysis of molecular contacts.



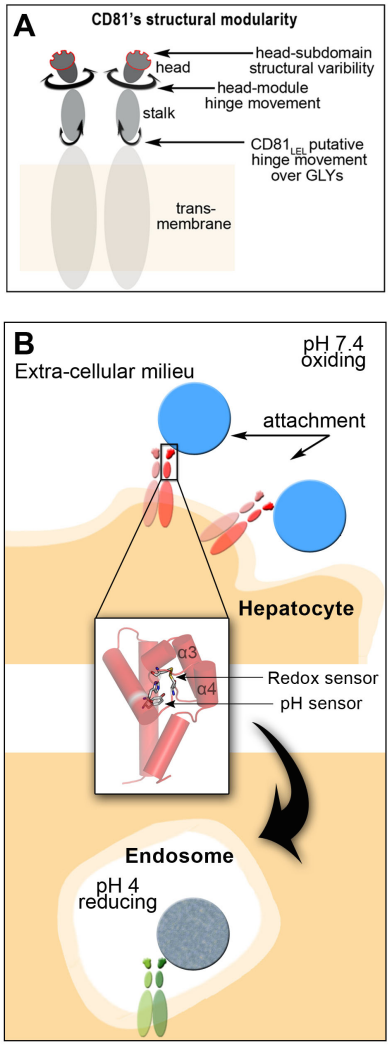
Difference in accessible surface area (Δ-ASA) by residue composing the head-subdomain for each molecule and for a given crystal form (from top-to-bottom same order as in Fig. 2A). Values within each panel correspond to the total Δ-ASA. Red-dots, green-dots and gold-dots identify molecules in *closed*, *open* and *intermediate* conformation, respectively. The read outline on the green-dot of mol-*12* indicates that α3 and a4 helices are almost parallel (θ = 134°).

Figure 4. Molecular dynamics of hCD81$_{LEL}$.



(A) Starting from an *open* conformation and a *closed* conformation (molecules *12* and *1*) we followed the RMSD evolution of the head-subdomain (blues) and the stalk-subdomain (greys). The head-subdomain shows most of the conformational variability. (B) Fluctuations of residues obtained after averaging 20 μs of MD trajectories. The moving elements of CD81$_{LEL}$ domain, colour coded in yellow, pink and red are mapped onto the structure. RMSF, root-mean-square-fluctuation. (C) Bidimensional distribution over the inter-helical angle (θ) and radius of gyration (R_g) space of the $2x10^5$ structures generated during MD simulations. Most visited conformers, regions labelled with 1, 2, 3 and 4 on the map and representative structures on the right, reproduce the solved structures; the colour scale on the far left, shows the probability of a sampled structure to belong to a given point in the conformational space. Histograms on top and right axes, representing the distribution of structural conformers, show that the closed conformations are overall favoured. (D) Influence of the pH on the CD81$_{LEL}$ conformers. Distributions of structures observed at pH=4.0 (red) and pH=7.4 (grey) using the R_g criteria for structural classification showing that the increasingly open conformations (R_g > 8 Å) a more favoured at acidic pH (see Movie S2). PDF is the probability density function. (E) Distribution of the S/S distance (between C157 and C175) derived from simulations at pH=7.4 (grey) and 4.0 (red) where the disulphide bridge was removed, starting from open (FFFF A), intermediate (EEEE A) and closed (AAAA A) conformations (all together). The grey curve shows a narrower distribution of S/S distances supporting a more stable configuration at pH = 7.4.

Figure 5. Putative CD81 mechanism of action and HCV cell-entry model



(A) Cartoon of the FL-CD81 structural modularity with the two hinge movements between head/stalk and stalk/transmembrane regions with the head-subdomain displaying structural variability.(B) Schematic model of HCV CD81 mediated cell entry. Top, HCV as light-blue circle binding at the extra-cellular milieu to the *closed* head-subdomain conformation of CD81 as red and light-red (multiple bindings could also occur); the complex then migrates to the tight junction where CD81 interacts with claudin-1 (not shown) pre-empting the uptake of the virus. Inset, the location of environmental sensing residues at the core of the $CD81_{LEL}$ structure; linked by the black-arrow the scenario in which, after internalization through the endocytotic pathway, the different endosomal conditions (acidic and redox environment) would favour the head-subdomain conformation to transit towards an *open* conformation with possible implications for the transition from a pre- to a post-fusion state of the HCV E1E2:CD81 complex.The oligomeric state of FL-CD81 has been depicted as dimer, in orange and light-orange the cell and the cell-membrane respectively (virus, receptor and cell are not to scale).

Table 1. $hCD81_{LEL}$ data collection and refinement statistics. The four novel crystal forms of $hCD81_{LEL}$ determined in this study (PDB IDs AAAA, BBBB, CCCC, DDDD), together with the higher resolution monoclinic (PDB ID EEEE) and hexagonal forms (PDB ID FFFF), the resolution of these crystal forms improves over the 1.6 Å PDB ID 1G8Q (13) and 2.6 Å PDB ID 1IV5 (14).

In parenthesis the values in the highest resolution shell.

(*) Data indexed, integrated and scaled in HKL2000 (24).

(¶) Data indexed and integrated with XDS and scaled with Aimless (25, 26).

(†) Obtained by merging two dataset collected from the same crystal.

All data were converted in SFs using Truncate (26)

PDB ID	**AAAA***	**BBBB***	**CCCC**[¶]	**DDDD**[*]	**EEEE***	**FFFF**[¶]
Crystallization conditions	29% EtOH, 0.092 M Citrate pH 2.2, 0.113 M Na_2HPO_4 pH 9.3, (pH ~5) 39% PEG 300	40% EtOH, 0.097 M Citrate pH 2.2, 0.113 M Na_2HPO_4 pH 9.3, (pH ~6) 2.5% PEG 1000	0.1 M NaCitrate pH 4.0, 0.8 M $(NH4)_2$ SO_4	0.3 M NaCl 0.1 Na Hepes, pH 7.5, 25% v/v PEG400 grown in presence of benzyl-salicylate	0.1 M MIB pH 5.0, 25% w/v PEG 1500	0.1 MMT pH 5, 25% w/v PEG 1500 grown in presence of fexofenadine
Data Collection Statistics						
Beamline	I03@Diamond	I03@Diamond	I02@Diamond	I02@Diamond	ID23-1@ESRF	I02@Diamond
Number of datasets (frames)	2 (720)[†]	1 (1200)	1 (720)	1 (720)	2 (2606)[†]	1 (500)
Oscillation angle	0.5°	0.1°	0.25°	0.25°	0.15°	0.2°
Unit cell (Å,°)	a=b=49.9, c=132.4	a=b=97.9, c=34.4	a=132.0, b=106.5, c=66.0, β= 99.4	a=97.3 b=137.8 c=129.7	a=31.5, b=75.9, c=37.5, β= 107.2	a=b=101.4 c=35.9
Space Group (Z)	$P6_222$ (XX)	$P3_1$ (12)	$C2$ (20)	$C222_1$ (32)	$P2_1$ (4)	$P6_4$ (12)
Resolution (Å)	44.2-1.90 (1.94-1.90)	49.0-2.38 (2.51-2.38)	53.1-3.10 (3.18-3.10)	50.3-5.00 (5.18-5.00)	26.0-1.28 (1.30-1.28)	50.7-2.02 (2.07-2.02)
Observations	944083	50600	51479	242760	1158652	57585
Unique observations	8341 (534)	14545 (2166)	15510 (1117)	3970 (394)	38270 (985)	12501 (566)
Rmerge	0.076 (-)	0.056 (0.715)	0.074 (0.525)	0.090 (-)	0.060 (0.385)	0.030 (0.396)
Rmeas	0.078 (-)	0.066 (0.839)	0.107 (0.726)	0.099 (-)	0.066 (0.448)	0.034 (0.483)
Rpim	0.018 (0.450)	0.034 (0.433)	0.058 (0.384)	0.046 (0.461)	0.029 (0.224)	0.014 (0.303)
CC1/2 (highest shell)	0.848	0.773	0.817	0.522	0.925	0.802
‹I/σ(I)›	50.3 (2.7)	12.3 (1.7)	10.1 (2.0)	17.1 (2.0)	43.0 (3.1)	29.1 (2.3)
Completeness %	100.0 (100.0)	98.1 (98.9)	98.7 (99.4)	99.9 (100.0)	88.6 (41.3)	88.3 (53.4)
Redundancy	30.5 (34.6)	3.5 (3.5)	3.3 (3.5)	6.1 (6.0)	5.3 (3.8)	4.6 (1.9)
Wilson- B factor	38.2	51.5	116.3	-	15.1	40.7
Refinement Statistics						
Resolution (Å)	1.90 (1.97-1.90)	2.38 (2.47-2.38)	3.1 (3.21-3.10)	5.0 (5.92-5.3)	1.28 (1.32-1.28)	2.02 (2.09-2.02)
Rwork	0.203 (0.307)	0.174 (0.308)	0.206 (0.292)	0.337	0.145 (0.201)	0.169 (0.252)
Rfree	0.259 (0.369)	0.227 (0.362)	0.249 (0.377)	0.351	0.177 (0.234)	0.216 (0.369)
Molecules/asu	1	4	5	4	2	2
Molecule number (Chain ID)	1 (A)	2 (A), 3 (B), 4 (C) , 5 (D)	6 (A), 7 (B), 8 (C), 9 (D), 10 (E)	(A), (B), (C), (D)	11 (A), 12 (B)	13 (A), 14 (B)
Rmsd bonds (Å)	0.009	0.010	0.007	-	0.010	0.006
Rmsd angles (º)	1.08	1.13	1.0	-	1.20	0.84
Ramachandran favoured (outliers)	97.0% (0)	95.0% (0)	93.0% (0)	-	97% (0)	96.0% (0)
Amino acids	88	359	432	-	178	182
Waters	10	10	0	-	163	17
Ligands	1 (PO_4)	12 (PO_4, EDO)	6 (SO_4, EDO)	-	4 (EDO)	5 (Cl, EDO)
$<B_{prot}>$ ($<B_{waters}>$; ($<B_{ligands}>$)(Å^2)	61.5 (58.1; 126.7)	71.6 (66.1; 92.9)	100.8 (-; 132.5)	-	23.3 (33.7; 55.2)	68.3 (58.6; 99.6)

.

References

1. Block TM, Mehta AS, Fimmel CJ, Jordan R. Molecular viral oncology of hepatocellular carcinoma. Oncogene 2003;22:5093-5107.
2. Rice JP. Hepatitis C treatment: Back to the warehouse. Clinical Liver Disease 2015;6:27-29.
3. Pileri P, Uematsu Y, Campagnoli S, Galli G, Falugi F, Petracca R, Weiner AJ, et al. Binding of hepatitis C virus to CD81. Science 1998;282:938-941.
4. Meredith LW, Wilson GK, Fletcher NF, McKeating JA. Hepatitis C virus entry: beyond receptors. Rev Med Virol 2012;22:182-193.
5. Higginbottom A, Quinn ER, Kuo CC, Flint M, Wilson LH, Bianchi E, Nicosia A, et al. Identification of amino acid residues in CD81 critical for interaction with hepatitis C virus envelope glycoprotein E2. J Virol 2000;74:3642-3649.
6. Kong L, Giang E, Nieusma T, Kadam RU, Cogburn KE, Hua Y, Dai X, et al. Hepatitis C virus E2 envelope glycoprotein core structure. Science 2013;342:1090-1094.
7. Yu X, Qiao M, Atanasov I, Hu Z, Kato T, Liang TJ, Zhou ZH. Cryo-electron microscopy and three-dimensional reconstructions of hepatitis C virus particles. Virology 2007;367:126-134.
8. Badia-Martinez D, Peralta B, Andres G, Guerra M, Gil-Carton D, Abrescia NG. Three-dimensional visualization of forming Hepatitis C virus-like particles by electron-tomography. Virology 2012;430:120-126.
9. Catanese MT, Uryu K, Kopp M, Edwards TJ, Andrus L, Rice WJ, Silvestry M, et al. Ultrastructural analysis of hepatitis C virus particles. Proc Natl Acad Sci U S A 2013;110:9505-9510.
10. Krey T, d'Alayer J, Kikuti CM, Saulnier A, Damier-Piolle L, Petitpas I, Johansson DX, et al. The disulfide bonds in glycoprotein E2 of hepatitis C virus reveal the tertiary organization of the molecule. PLoS Pathog 2010;6:e1000762.
11. Khan AG, Whidby J, Miller MT, Scarborough H, Zatorski AV, Cygan A, Price AA, et al. Structure of the core ectodomain of the hepatitis C virus envelope glycoprotein 2. Nature 2014.
12. El Omari K, Iourin O, Kadlec J, Sutton G, Harlos K, Grimes JM, Stuart DI. Unexpected structure for the N-terminal domain of hepatitis C virus envelope glycoprotein E1. Nat Commun 2014;5:4874.
13. Kitadokoro K, Bordo D, Galli G, Petracca R, Falugi F, Abrignani S, Grandi G, et al. CD81 extracellular domain 3D structure: insight into the tetraspanin superfamily structural motifs. EMBO J 2001;20:12-18.
14. Kitadokoro K, Ponassi M, Galli G, Petracca R, Falugi F, Grandi G, Bolognesi M. Subunit association and conformational flexibility in the head subdomain of human CD81 large extracellular loop. Biol Chem 2002;383:1447-1452.
15. Seigneuret M. Complete predicted three-dimensional structure of the facilitator transmembrane protein and hepatitis C virus receptor CD81: conserved and variable structural domains in the tetraspanin superfamily. Biophys J 2006;90:212-227.
16. Olaby RA, Azzazy HM, Harris R, Chromy B, Vielmetter J, Balhorn R. Identification of ligands that target the HCV-E2 binding site on CD81. J Comput Aided Mol Des 2013;27:337-346.
17. Holzer M, Ziegler S, Neugebauer A, Kronenberger B, Klein CD, Hartmann RW. Structural modifications of salicylates: inhibitors of human CD81-receptor HCV-E2 interaction. Arch Pharm (Weinheim) 2008;341:478-484.
18. Neugebauer A, Klein CD, Hartmann RW. Protein-dynamics of the putative HCV receptor CD81 large extracellular loop. Bioorg Med Chem Lett 2004;14:1765-1769.
19. Rajesh S, Sridhar P, Tews BA, Feneant L, Cocquerel L, Ward DG, Berditchevski F, et al. Structural basis of ligand interactions of the large extracellular domain of tetraspanin CD81. J Virol 2012;86:9606-9616.
20. Sharma NR, Mateu G, Dreux M, Grakoui A, Cosset FL, Melikyan GB. Hepatitis C virus is primed by CD81 protein for low pH-dependent fusion. J Biol Chem 2011;286:30361-30376.
21. Harman C, Zhong L, Ma L, Liu P, Deng L, Zhao Z, Yan H, et al. A view of the E2-CD81 interface at the binding site of a neutralizing antibody against hepatitis C virus. J Virol 2015;89:492-501.
22. Drummer HE, Wilson KA, Poumbourios P. Identification of the hepatitis C virus E2 glycoprotein binding site on the large extracellular loop of CD81. J Virol 2002;76:11143-11147.
23. Aricescu AR, Lu W, Jones EY. A time- and cost-efficient system for high-level protein production in mammalian cells. Acta Crystallogr D Biol Crystallogr 2006;62:1243-1250.
24. Otwinoski ZaM, W.: Processing of X-ray Diffraction Data Collected in Oscillation Mode. In: Carter CWS, R. M., ed. Methods in Enzymology. Volume 276. New York: Academic Press, 1997; 307-326.
25. Vonrhein C, Flensburg C, Keller P, Sharff A, Smart O, Paciorek W, Womack T, et al. Data processing and analysis with the autoPROC toolbox. Acta Crystallogr D Biol Crystallogr 2011;67:293-302.
26. Collaborative Computational Project N. The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr 1994;50:760-763.
27. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 2010;66:213-221.
28. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr 2010;66:486-501.
29. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 2004;32:W500-502.
30. Lee HS, Choi J, Yoon S. QHELIX: a computational tool for the improved measurement of inter-helical angles in proteins. Protein J 2007;26:556-561.
31. Rostkowski M, Olsson MH, Sondergaard CR, Jensen JH. Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. BMC Struct Biol 2011;11:6.
32. Sfriso P, Hospital A, Emperador A, Orozco M. Exploration of conformational transition pathways from coarse-grained simulations. Bioinformatics 2013;29:1980-1986.
33. Meyer T, D'Abramo M, Hospital A, Rueda M, Ferrer-Costa C, Perez A, Carrillo O, et al. MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. Structure 2010;18:1399-1409.
34. Hospital A, Andrio P, Fenollosa C, Cicin-Sain D, Orozco M, Gelpi JL. MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. Bioinformatics 2012;28:1278-1279.
35. D.A. Case VB, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo,

B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman. AMBER 14. University of California, San Francisco 2014.
36. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 2010;78:1950-1958.
37. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins 2005;61:704-721.
38. Gelpi JL, Kalko SG, Barril X, Cirera J, de La Cruz X, Luque FJ, Orozco M. Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. Proteins 2001;45:428-437.
39. Stuart DI, Levine M, Muirhead H, Stammers DK. Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 A. J Mol Biol 1979;134:109-142.
40. Eyal E, Gerzon S, Potapov V, Edelman M, Sobolev V. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J Mol Biol 2005;351:431-442.
41. Orozco M. A theoretical view of protein dynamics. Chem Soc Rev 2014;43:5051-5066.
42. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, et al. Atomic-level characterization of the structural dynamics of proteins. Science 2010;330:341-346.
43. Shan Y, Gnanasambandan K, Ungureanu D, Kim ET, Hammaren H, Yamashita K, Silvennoinen O, et al. Molecular basis for pseudokinase-dependent autoinhibition of JAK2 tyrosine kinase. Nat Struct Mol Biol 2014;21:579-584.
44. Dalmas O, Sompornpisut P, Bezanilla F, Perozo E. Molecular mechanism of Mg2+-dependent gating in CorA. Nat Commun 2014;5:3590.
45. Anderson AC. The process of structure-based drug design. Chem Biol 2003;10:787-797.
46. Stroupe C, Brunger AT. Crystal structures of a Rab protein in its inactive and active conformations. J Mol Biol 2000;304:585-598.
47. Lange CM, Jacobson IM, Rice CM, Zeuzem S. Emerging therapies for the treatment of hepatitis C. EMBO Mol Med 2014;6:4-15.
48. Evans MJ, von Hahn T, Tscherne DM, Syder AJ, Panis M, Wolk B, Hatziioannou T, et al. Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry. Nature 2007;446:801-805.
49. Gerold G, Meissner F, Bruening J, Welsch K, Perin PM, Baumert TF, Vondran FW, et al. Quantitative Proteomics Identifies Serum Response Factor Binding Protein 1 as a Host Factor for Hepatitis C Virus Entry. Cell Rep 2015;12:864-878.
50. El Omari K, Iourin O, Harlos K, Grimes JM, Stuart DI. Structure of a pestivirus envelope glycoprotein E2 clarifies its role in cell entry. Cell Rep 2013;3:30-35.

7. 2 Efficient Relaxation of Protein−Protein Interfaces by Discrete Molecular Dynamics Simulations

Context:

Proteins are social molecules. They need to interact with their surrounding molecules to perform their function and especially other proteins. Predicting the interaction geometry of two known proteins is an extremely challenging task due to the large number of degrees of freedom in the system. The most successful approach, the rigid protein-protein docking, divides the problem in two parts: first, it essays all possible approximate orientations treating both proteins as rigid blocks then evaluating the energy of each binding pose. Rigid docking aims for computational efficiency but lacks of sensitivity when the interacting proteins deform while binding. Here, we present an automated method to consider protein flexibility in rigid docking poses. The protocol runs short dMD simulations over the best-ranked protein rigid complexes allowing the protein interface to adapt to its partner. The method is able to systematically improve the accuracy in the prediction protein-protein interaction pose, being remarkable for complexes with large conformational changes upon binding.

Title: Efficient Relaxation of Protein–Protein Interfaces by Discrete Molecular Dynamics Simulations

Authors: Agustí Emperador, Albert Solernou, Pedro Sfriso, Carles Pons, Josep Lluis Gelpí, Juan Fernandez Recio and Modesto Orozco
Stage: Published
Journal: Journal Chemical Theory and Computation
Type: Research Article
Supplementary Material: http://pubs.acs.org/doi/abs/10.1021/ct301039e

Author Contribution: PS designed the scoring function and energy potentials, performed some simulations, and contributed to the writing of the paper.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Efficient Relaxation of Protein−Protein Interfaces by Discrete Molecular Dynamics Simulations

Agusti Emperador,[†,§] Albert Solernou,[‡,§] Pedro Sfriso,[†,§] Carles Pons,[‡,§] Josep Lluis Gelpi,[‡,§,∥] Juan Fernandez-Recio,*[,‡,§] and Modesto Orozco*[,†,‡,§,∥]

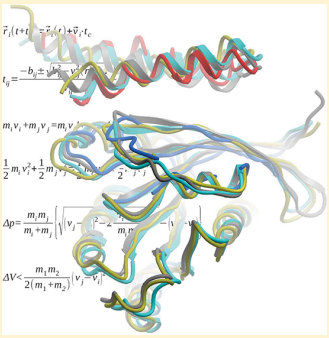
[†]Institute for Research in Biomedicine (IRB Barcelona), Baldiri i Reixac 10, Barcelona 08028, Spain

[‡]Barcelona Supercomputing Center, Jordi Girona 29, Barcelona 08034, Spain

[§]Joint BSC-IRB Research Program in Computational Biology, Barcelona, Spain

[∥]Departament de Bioquímica, Facultat de Biologia, Avgda Diagonal 645, Barcelona 08028, Spain

ABSTRACT: Protein−protein interactions are responsible for the transfer of information inside the cell and represent one of the most interesting research fields in structural biology. Unfortunately, after decades of intense research, experimental approaches still have difficulties in providing 3D structures for the hundreds of thousands of interactions formed between the different proteins in a living organism. The use of theoretical approaches like docking aims to complement experimental efforts to represent the structure of the protein interactome. However, we cannot ignore that current methods have limitations due to problems of sampling of the protein−protein conformational space and the lack of accuracy of available force fields. Cases that are especially difficult for prediction are those in which complex formation implies a non-negligible change in the conformation of the interacting proteins, i.e., those cases where protein flexibility plays a key role in protein−protein docking. In this work, we present a new approach to treat flexibility in docking by global structural relaxation based on ultrafast discrete molecular dynamics. On a standard benchmark of protein complexes, the method provides a general improvement over the results obtained by rigid docking. The method is especially efficient in cases with large conformational changes upon binding, in which structure relaxation with discrete molecular dynamics leads to a predictive success rate double that obtained with state-of-the-art rigid-body docking.



INTRODUCTION

Proteins are social molecules that exert their biological action through the interaction with other molecules. In particular, most signal transduction mechanisms in the cell are mediated by protein−protein interactions, which define complex interaction networks. Massive proteomics studies outline the components of different complexes in the cell, providing the first pictures of cellular interactomes. This provides knowledge of which molecular partners are participating in a given protein−protein interaction, but in order to understand its function at the molecular level or to interfere in the complex formation by small compounds, detailed information on the three-dimensional structure of the complexes is required.

Atomic resolution experimental techniques, especially X-ray crystallography and NMR, are providing an increasing amount of information on the three-dimensional structures of protein−protein complexes. However, the number of deposited complex structures in the Protein Data Bank[1] is currently less than 10% of the number of known binary interactions between human proteins[2] and a minuscule fraction of the estimated number of total interactions including transient complexes.[3] This has encouraged the use of computational methods, especially protein docking algorithms, which in the absence of atomic resolution structural data can provide useful information on protein complexes in the context of systems and network biology.

Many different protein−protein docking procedures have been reported to provide reduced sets of docking poses ideally enriched in near-native conformations and selected out of thousands or millions of alternative poses.[4,5] To achieve computational efficiency, and to reduce false positives, protein monomers are considered in most cases as rigid entities during the process. This rigid-body approach gives excellent results for those cases in which proteins show very little flexibility upon binding (e.g., average receptor and ligand unbound−bound backbone RMSD < 0.5 Å)[6] and therefore seem to associate following the "lock-and-key" mechanism.[7] However, for the majority of cases, complex formation involves larger conformational rearrangement,[8] and therefore the rigid docking approach yields poorer results.[6] Interestingly, the problematic cases for docking can be reasonably predicted based only on the intrinsic flexibility of the unbound state of the protein,[9] so treatment of flexibility is currently one of the major challenges in protein docking.

Different techniques to account for protein flexibility, typically at later stages of the docking procedure, have been developed.[10,11] In one of the first refinement methods, side-chain conformational optimization showed improvement in the docking predictions in specific cases, e.g., in those with a few

Received: July 23, 2012
Published: December 17, 2012

Table 1. Docking Results before and after DMD Relaxation

PDB	interface RMSD[a]	pyDock rank[b]	DMD rank[b]	PDB	interface RMSD[a]	pyDock rank[b]	DMD rank[b]
1AY7	0.54	17	10	1R6Q	1.67	40	14
1AZS	0.72	12	81	1RV6	1.09	6	2
1B6C	1.96	1	1	1T6B	0.62	56	23
1BUH	0.75	65	51	1TMQ	0.86	1	44
1BVK	1.24	52	69	1UDI	0.9	1	4
1BVN	0.87	2	8	1XD3	1.24	2	5
1CLV	0.86	1	1	1XQS	1.77	13	50
1E6E	1.33	3	10	1XU1	1.3	18	75
1E6J	1.05	35	20	1Z0K	0.53	7	1
1E96	0.71	1	1	1Z5Y	1.23	79	12
1EWY	0.8	28	66	1ZHI	0.68	2	5
1F51	0.74	36	59	2ABZ	0.9	18	8
1FFW	1.43	74	25	2AYO	1.39	24	72
1FLE	1.02	3	24	2B42	0.72	1	1
1FSK	0.45	3	16	2BTF	0.75	33	39
1GLA	0.98	50	51	2FD6	1.07	17	2
1GPW	0.65	1	23	2G77	1.08	13	5
1H9D	1.32	26	51	2HLE	1.4	2	2
1IQD	0.48	8	1	2HQS	1.14	15	18
1J2J	0.63	19	11	2HRK	2.03	16	8
1JTG	0.49	1	2	2I25	1.21	40	76
1K74	0.8	14	6	2JEL	0.17	42	34
1KKL	2.2	81	3	2O8V	1.37	60	84
1M10	2.1	81	2	2PCC	0.39	49	83
1MAH	0.61	19	8	2SIC	0.36	6	1
1N8O	0.94	53	35	2SNI	0.35	3	1
1NW9	1.97	30	4	2VDB	0.47	5	6
1OPH	1.21	14	1	3SGQ	0.39	98	36
1OYV	0.47	84	95	4CPA	0.62	10	14
1PPE	0.44	6	2	7CEI	0.7	11	1
1R0R	0.45	3	64				

[a]Interface C_α RMSD between unbound and bound molecules (in Å). [b]Best rank of any near-native docking conformation.

problematic interface residues,[12,13] but its applicability was found to be more limited for difficult cases with large interfaces.[14] Good predictions were also obtained with backbone refinement when experimental restraints were used, as evaluated on a limited number of test cases.[15] Another approach based on the deformation of the proteins along the low-energy normal modes[16] was applied to a test set of 10 complexes with a significant difference between unbound and bound structures, giving good results for some of the complexes.[17] Backbone flexibility was also included via Monte Carlo refinement[18] based on the Rosetta method.[19] This procedure was applied to a choice of 25 out of 49 complexes of the first Weng benchmark[20] with successful results. The above methods, typically quite expensive from a computational point of view, have shown overall good results as part of a general docking strategy in the international CAPRI assessment of docking predictions (http://www.ebi.ac.uk/msd-srv/capri/), but more efforts need to be done to assess the real improvement of refinement with respect to rigid-body docking in large sets of complexes.

In this work, we present a new method to treat protein flexibility in docking computations using discrete molecular dynamics (DMD).[21−25] DMD is very efficient from a computational point of view and allows a very simple change of granularity, from atomistic to coarse grained levels and a complete control on the sampling space, from the side-chain only to the entire protein. The technique, which has been used with success to study protein aggregation[26,27] and conformational transitions,[28,29] assumes that particles move at constant velocity (ballistic regime) from collision to collision, avoiding then the time-consuming integration of Newton's equations of motion every femtosecond and reducing dramatically the cost of calculations, without a dramatic loss of accuracy with respect to standard atomistic simulation techniques.[30,31] The method developed here has been tested with very good success on Weng's protein−protein docking benchmark 4.0,[32] for which the structures of both the complex and the unbound partners are experimentally known.

OUTLINE OF THE METHOD

Protein Benchmark. We have used Weng's protein docking benchmark 4.0,[32] with known X-ray structures for both the unbound and the bound subunits. For test purposes, we selected only the 61 cases (Table 1) in which rigid-body docking with pyDock finds at least one near-native docking orientation (see below) among the 100 top-ranked conformations. Near-native docking solutions were defined (according to one of the criteria used in CAPRI) as those with interface RMSD < 4 Å RMSD with respect to the reference complex structure, being the interface RMSD calculated for the C_α atoms in such an interface. The success rates of the predictions are defined as the percentage of test cases in which at least one near-native docking solution was found within the top N solutions. We classified the benchmark cases according to the flexibility upon binding, for which we used the interface C_α RMSD between

bound and unbound molecules, as defined in the protein benchmark 4.0.[32]

Rigid-Body Sampling and Scoring. We generated a set of 10 000 docking models for each protein−protein complex by using FTDOCK,[33] based on geometrical complementarity, using a grid resolution of 0.7 Å for the representation of proteins, and including an electrostatics term. These models were evaluated using pyDock,[34] a physics-based docking scoring function with excellent results in standard benchmarks as well as in CAPRI.[35,36] PyDock calculates the binding energy between the interacting proteins based on (i) a truncated and linearly screened electrostatic term, (ii) a truncated and weighted van der Waals term, and (iii) an accessible surface area (ASA)-based desolvation energy term with atomic parameters previously optimized for protein docking.[37] We selected for DMD relaxation the subset of the 100 best scored poses according to pyDock.

General Discrete Molecular Dynamics (DMD) Formalism. DMD is based on the use of discontinuous, stepwise potentials, which guarantee that particles move in the ballistic regime, following a linear movement with constant velocity until they reach a potential step:[22]

$$\vec{r}_i(t + t_c) = \vec{r}_i(t) + \vec{v}_i \cdot t_c \tag{1}$$

where and $\vec{r}_i$ and $\vec{v}_i$ stand for positions and velocities and t_c is the minimum among the collision times t_{ij} between each pair of particles i and j:

$$t_{ij} = \frac{-b_{ij} \pm \sqrt{b_{ij}^2 - v_{ij}^2(r_{ij}^2 - d^2)}}{v_{ij}^2} \tag{2}$$

where r_{ij} is the magnitude of $\vec{r}_{ij} = \vec{r}_j - \vec{r}_i$, v_{ij} is the magnitude of $\vec{v}_{ij} = \vec{v}_j - \vec{v}_i$, $b_{ij} = \vec{r}_{ij}\,\vec{v}_{ij}$, and d is the distance corresponding to the wall of the square well.

The collision happens when the particle distance is that corresponding to a potential step. Then there is a transfer of linear momentum into the direction of the vector $\vec{r}_{ij}$:

$$m_i\vec{v}_i = m_i\vec{v}_i' + \Delta\vec{p}$$

$$m_j\vec{v}_j + \Delta\vec{p} = m_j\vec{v}_j' \tag{3}$$

where the prime indices denote the velocities after the collision.

The changes in velocities upon collision are derived by applying conservation of energy and momentum:

$$m_iv_i + m_jv_j = m_iv_i' + m_jv_j' \tag{4}$$

$$\frac{1}{2}m_iv_i^2 + \frac{1}{2}m_jv_j^2 = \frac{1}{2}m_iv_i'^2 + \frac{1}{2}m_jv_j'^2 + \Delta V \tag{5}$$

where ΔV stands for the size of the potential energy step.

The transferred momentum can be easily determined from

$$\Delta p = \frac{m_im_j}{m_i + m_j}\left\{\sqrt{(v_j - v_i)^2 - 2\frac{m_i + m_j}{m_im_j}\Delta V} - (v_j - v_i)\right\} \tag{6}$$

Note that, in case that $\Delta V > 0$, the two particles can overcome the step as long as

$$\Delta V < \frac{m_1m_2}{2(m_1 + m_2)}(v_j - v_i)^2 \tag{7}$$

Otherwise, if the particles remain in the potential well, eq 6 reduces to

$$\Delta p = \frac{m_im_j}{m_i + m_j}\{\sqrt{(v_j - v_i)^2} - (v_j - v_i)\} \tag{8}$$

which taking the negative solution of the root leads to

$$\Delta p = \frac{2m_im_j}{m_i + m_j}(v_i - v_j) \tag{9}$$

To obtain good computational performance, we used here a simple implementation of DMD that uses a force-field containing "bonded" and "non-bonded" terms. The first ones include stretchings, bendings, and torsions (all represented by bond or pseudobond lengths) involving double or conjugated bonds, which are represented by means of infinite square wells with a width (derived from analysis of large database of atomistic MD simulations[22]) equal to 1% of the bond/pseudobond distance.

Regarding the nonbonded part, our force field includes solvation, van der Waals, and Coulomb electrostatic terms

$$V = V_{solv} + V_{VdW} + V_{Coul} \tag{10}$$

The van der Waals (V_{vdW}) and Coulomb (V_{Coul}) terms are the DMD version of the Lennard-Jones and Coulomb potentials, while solvation (V_{solv}) was accounted by the DMD version[22] of the Lazaridis and Karplus EEF1 effective energy function.[38] Intramolecular hydrogen bonds were restrained by square wells which guarantee the maintenance of secondary structure elements during the DMD simulations.

DMD Implementation for Protein−Protein Interactions. To increase computational efficiency, we used a multiscale representation of the proteins, where the level of detail and the allowed flexibility of the residues decrease as the distance to the protein interface increases. Thus, residues in the protein−protein interface (residues with at least one atom less than 8 Å from any atom of the other protein in the rigid docking pose) were defined at the all-heavy-atoms level, keeping them completely flexible. A second layer was defined by residues located at 8−12 Å from any atom of the other protein, where all-heavy-atoms representation was used, but atom positions were restrained by Go-like square wells. The rest of the protein was represented at the C_α level, using Go-like square wells to restrain their movements. DMD trajectories were long enough to ensure equilibration of the docking conformations (see Results and Discussion), and we used as the scoring function the average of the potential energy over the last 15% of the trajectory for each conformation.

Analysis of Native Contacts along Simulation. We considered that two residues defined a native contact if any of their atoms were interacting in the experimental complex structure, i.e., if one atom from the first residue and another atom from the other residue were within the interaction distance defined in the DMD potentials (that were stepwise functions with a finite range; see above).

Statistical Significance of DMD Improvement in Flexible Cases. We have applied a Wilcoxon rank-sum test to evaluate whether improvement of the results after DMD relaxation with respect to rigid-body docking is more evident for flexible cases. We built two groups of cases, one formed by 12 cases in which the DMD relaxation *significantly improved* with respect to rigid-body docking (best near-native rank went from >10 to <10) and the other one formed by six cases in which DMD relaxation *significantly worsened* with respect to rigid-body docking (best near-native rank went from <10 to >10).

We considered that having or not a near native within the top 10 docking solutions is a reasonable criterion in order to evaluate the success of a docking result (as used in CAPRI). For higher ranks, a small improvement or worsening is irrelevant for performance assessment purposes. The Wilcoxon rank-sum test proved that the distributions of unbound−bound RMSD values in these two groups of cases differed significantly (Mann−Whitney $U = 13$, $n_1 = 12$, $n_2 = 6$, $p < 0.05$ two-tailed).

Estimating Binding Flexibility from Normal Modes. The extent of the conformational motion of a protein due to thermal fluctuations can be estimated within the normal-mode analysis (NMA) framework.[16] NMA assumes that the protein structure is a system of coupled harmonic oscillators connecting the C_α atoms. The dynamics of such a system is constituted by the normal modes, each one involving all the particles of the system. We estimated the total deformation of a protein as the average of the amplitudes due to each mode (eq 11):

$$\mathrm{RMSD}_{\mathrm{predicted}} = \sqrt{\frac{1}{N}\sum \lambda} \tag{11}$$

where the sum runs over all the normal modes $(3N - 6)$, and N is the number of C_α atoms. At a given temperature, the amplitude λ of the motion due to a normal mode is estimated as in eq 12:[39]

$$\lambda = \frac{k_B T}{k} \tag{12}$$

where k is the stiffness associated to the mode; therefore the relevant conformational changes will be produced by the softest normal modes (those with the lowest stiffness). We have assigned the predicted deformation of each complex as the average value of the predicted deformation of ligand and receptor. To classify the complexes as (allegedly) rigid and flexible, we have used a $\mathrm{RMSD}_{\mathrm{predicted}}$ cutoff value of 0.43 Å, which is the average of the predicted deformations over the complete benchmark. With this cutoff value, 60% of the complexes in the benchmark were classified as flexible. This coincides with the fraction that is considered flexible when taking the experimental interface deformations upon binding (using 1 Å as the interface RMSD cutoff value), but some complexes were classified in different groups when using each criterion.

RESULTS AND DISCUSSION

DMD Improves Docking Predictions. We performed rigid docking using our pyDock approach on the cases of Weng's protein docking benchmark 4.0,[32] for which both bound and unbound conformations of the interacting proteins are experimentally known. After a sequential scoring procedure (see Outline of the Method), we selected the 61 complexes for which the rigid docking procedure provided at least one near-native conformation within the 100 top scored docking poses (we limited the number of analyzed docking poses to 100 in order to perform a systematic benchmark in a reasonable time). This yielded a total of 6100 docking poses that were subjected to multiscale DMD simulations. It is worth noting that contrary to other protocols, such as atomistic MD simulation, which would demand simulation times orders of magnitude higher,[40] our DMD procedure allowed us to complete the process in less than 10 h on our 512 core Xeon Cluster (6100 trajectories for 61 complexes), making the method compatible with high-throughput procedures.

We found that DMD generally improves the docking energy landscapes, decreasing more effectively the energy of docking poses that are closer to the crystallographic complex structure. This is for example the case of 1Z0K, a complex that undergoes a small conformational change upon binding (0.53 Å interface RMSD between unbound and bound forms). Figure 1A shows

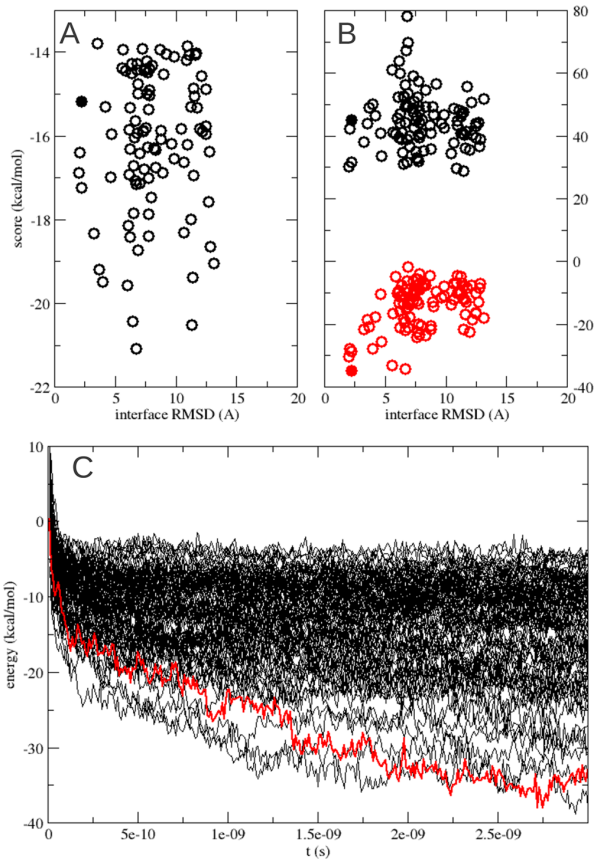


Figure 1. Rigid-body docking and DMD relaxation for complex 1Z0K. (A) pyDock score vs interface RMSD for each rigid-body docking conformation of the complex 1Z0K is shown. (B) DMD score vs interface RMSD for each conformation, before the simulations (rigid-body docking structures; in black) and after the DMD relaxation (in red). Filled circles represent the best-scoring near-native solution after DMD relaxation, shown also before DMD for comparison. (C) Evolution of the energies of the 100 docking conformations for the 1Z0K complex during the DMD simulations. The first ranked near-native conformation after relaxation has been highlighted in red.

the pyDock scoring of the rigid-body docking poses distributed according to interface RMSD with respect to the crystallographic complex structure. Figure 1B shows the distribution of these docking solutions according to DMD scoring before and after 3 ns DMD relaxation. Remarkably, the rank 1 docking solution after DMD relaxation (filled circle) is one of the best near-native rigid docking poses before relaxation in structural terms (2.2 Å interface RMSD from the experimental structure) but is not so good in terms of rigid-body energy (rank 67 and 57 by pyDock and DMD scoring before relaxation, respectively). The effect of DMD relaxation in docking refinement is clear in Figure 1C, which shows the evolution of the DMD potential energies of each docking pose for complex 1Z0K during the

simulation, with the red line corresponding to the rank 1 solution after DMD relaxation. We found that DMD trajectories of 3 ns were sufficient to reach equilibrium. Clearly, the DMD procedure leads to a decrease in the energy of all conformations, but near-native conformations experience a deeper energy improvement, populating the best-scoring docking poses after DMD relaxation. In terms of predictive results, this case was not that bad for rigid-body pyDock, since there was another near-native solution with rank 7 (Table 1), albeit with higher RMSD (3.7 Å), but DMD relaxation was more efficient in the near-native solutions closer to the bound state and yielded an overall improvement in the docking energy landscapes and in the predictive results.

The results of the described docking and DMD procedure on the entire data set of complexes (the largest benchmark available) demonstrate that DMD relaxation significantly improves the success rate of the docking predictions, by systematically improving the ranking of the near-native docking poses. For instance, the success rate for the top 10 scoring structures (which is a reasonable number of docking models in a realistic scenario, e.g, the number of models submitted to CAPRI) increased from 39% for the rigid-body procedure to 50% after 3 ns DMD relaxation (Figure 2A). Interestingly, the improvement in success rates was already evident from the very beginning of the DMD trajectory. After only 200 ps, the general results were very similar to those after 3 ns, suggesting that the fast side-chain movements have a fundamental influence on the improvement in energy of the docking conformations (as discussed below).

Improvement of Predictions after DMD Relaxation Is More Evident for Flexible Cases. In order to check which types of complexes benefit the most from DMD relaxation, we reanalyzed our results by grouping cases depending on the extent of the binding-induced geometrical changes upon interaction. Results summarized in Figure 2C indicate that 3 ns DMD leads to a very significant improvement in the ranking of docking poses in the case of complexes showing large conformational changes upon binding (interface deformation above 1.0 Å RMSD). For these cases, the success rates obtained with DMD for the top 1 and top 10 docking poses were 8% and 48%, respectively, i.e., twice those obtained with rigid-body docking. Among these flexible cases, there were actually seven cases in which the results significantly improved (best near-native rank went from >10 to <10 after DMD relaxation), while only one successful docking case became significantly worse after DMD relaxation. In the case of complexes with low conformational change upon binding, the success rates obtained with DMD

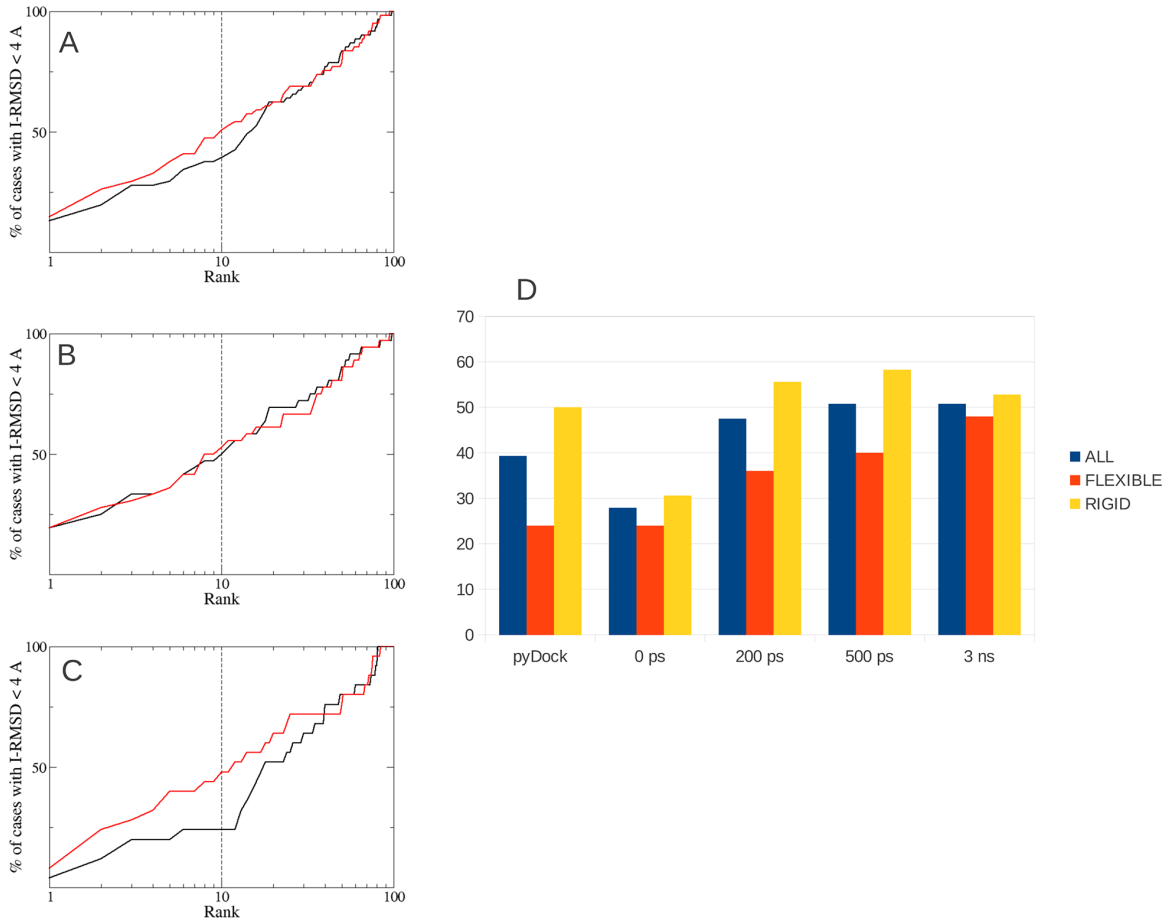


Figure 2. Docking success rates computed for the data set of protein−protein complexes used here. Black line, rigid-body docking ranked with the pyDock scoring function; red lines, docking structures relaxed after 3 ns DMD simulations. (A) Success rates for the entire data set. (B) Success rates for complexes with small conformational changes upon binding (interface RMSD unbound−bound <1 Å). (C) Success rates for complexes with large conformational changes upon binding (interface RMSD unbound−bound >1 Å). (D) Success rates for top 10 docking solutions at different times of the simulation, for rigid and flexible cases.

were similarly high to those obtained with rigid docking (Figure 2B). There were five cases that significantly improved and another five that significantly worsened. A Wilcoxon rank-sum test proved that the improvement in the results after DMD relaxation is significantly more evident for flexible cases (see Outline of the Method). This is in an important result, since a flexible refinement should not only improve the results in flexible cases but also not degrade the quality of docking models where binding-induced conformational changes are negligible. It has been previously reported that optimization of non-native interfaces may increase the number of false positives,[12] which does not seem to happen here.

The pace of improvement of the success rates along the DMD simulation is also different for the rigid and flexible cases. While for the rigid cases (Figure 2B) the optimal predictive results were achieved from the very beginning (at 200 ps, success rates are very similar to those after 3 ns), for the flexible cases (Figure 2C) the improvement of the predictive rates with respect to rigid-body docking is more progressive and actually shows the best results at 3 ns. For the purpose of clarity, Figure 2D shows the evolution of the top 10 success rates for rigid and flexible cases at different stages of the DMD trajectories. This confirms that DMD relaxation does not further improve the already good docking success rates in rigid cases, while flexible cases need some conformational rearrangement before improving docking success rates to achieve values more similar to the rigid cases. We will explore in the next section the nature of these conformational rearrangements.

Structural and Energetic Determinants of DMD Docking Success. The DMD relaxation procedure leads to non-negligible changes in the complex conformations obtained by rigid-body docking, with extensive modification of the whole interface (involving not only side chain but also backbone movements). In many cases, the deformation of both ligand and receptor along DMD relaxation, as it occurs in any standard molecular dynamics simulations,[40] exceeds the typical conformational changes upon binding (around 1−2 Å RMSD). In some successful cases, like 1Z0K, the relaxation makes the majority of the interface of a docking solution get closer to the crystallographic complex structure, which can explain the improvement in the predictions (Figure 3A). The conformational changes are shown to be similar to those we found making explicit solvent MD simulations with GROMACS. Indeed, the best near-native solution in this case conserved the same number of native contacts during DMD relaxation as the native complex (Figure 4A), which made this solution achieve the same energy as the native complex (Figure 4B). However, in other successful cases like 1M10, DMD relaxation cannot reproduce the whole native interface (Figure 3B; Figure 4D). In the past, it has been reported that refinement can be beneficial just by reorientation of the side chains to adopt conformations that improve the binding energy, even though overall interface conformation does not necessarily get closer to that in the native complex.[15] This seems to be the case here. We can explain the reasons for DMD improvement, in spite of the interface deformation, in terms of the type of side-chain contacts formed during simulation. Figure 4F shows the evolution of the native contacts that form attractive interactions during simulation for different docking solutions in the 1M10 case. The near-native solution shows a rapid formation of a few attractive native contacts that are ultimately responsible for its favorable energy. Although its energy value is not as good as that of the native complex (Figure 4E), it is sufficient to improve the rank of this near-native solution from 81 to 2 after DMD. Simultaneously, the number of repulsive contacts dramatically reduces in the first stages of the DMD relaxation (data not shown).

Overall, these results indicate that during the DMD simulation the structures relax to reduce the repulsive contacts that arose from incorrect rigid unbound conformations (between regions with electric charges of the same sign, steric clashes, unfavorable desolvation, etc.) and to increase the attractive native interactions found in the experimental complex structures (hydrophobic, salt bridges, etc.). This effect should be especially important in the flexible cases, in which DMD relaxation represents a greater advantage with respect to rigid approaches. As a consequence, the relaxation method is very efficient in improving the energy of many of the near-native conformations, which is reflected in the improvement of the docking predictions.

Can We Identify a Priori the Cases That Will Benefit from DMD Relaxation? We showed that proteins that undergo large conformational movements upon binding are

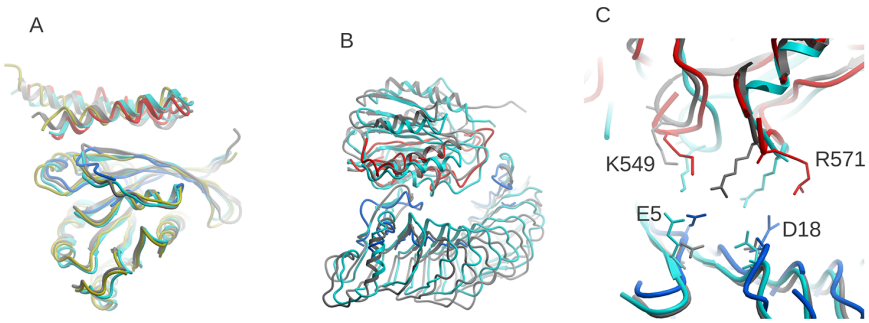


Figure 3. Conformational changes in near-native docking conformations after DMD relaxation. (A) Rank 1 near-native docking conformation of the complex 1Z0K after DMD relaxation (receptor in blue, ligand in red). The rigid docking conformation (initial structure) is shown in gray. For comparison, the experimental complex structure is shown in light blue, and the relaxed structure after simulation with GROMACS is shown in yellow. The relaxed protein−protein interface after DMD is getting closer to the experimental complex interface. (B) Rank 1 near-native docking conformation of the complex 1M10 after DMD relaxation (same color code as in A). In spite of the improvement in ranking, the relaxed protein−protein interface is not closer to the complex structure. (C) Key residue−residue native contacts for the 1M10 complex are salt-bridges between glycoprotein IB-α E5 and Von Willebrand Factor K549, and between D18 and R571 (in cyan). These were formed in the rank 2 conformation after DMD relaxation (receptor in blue, ligand in red), but they were not formed in the rigid-body docking solution before relaxation (in gray).

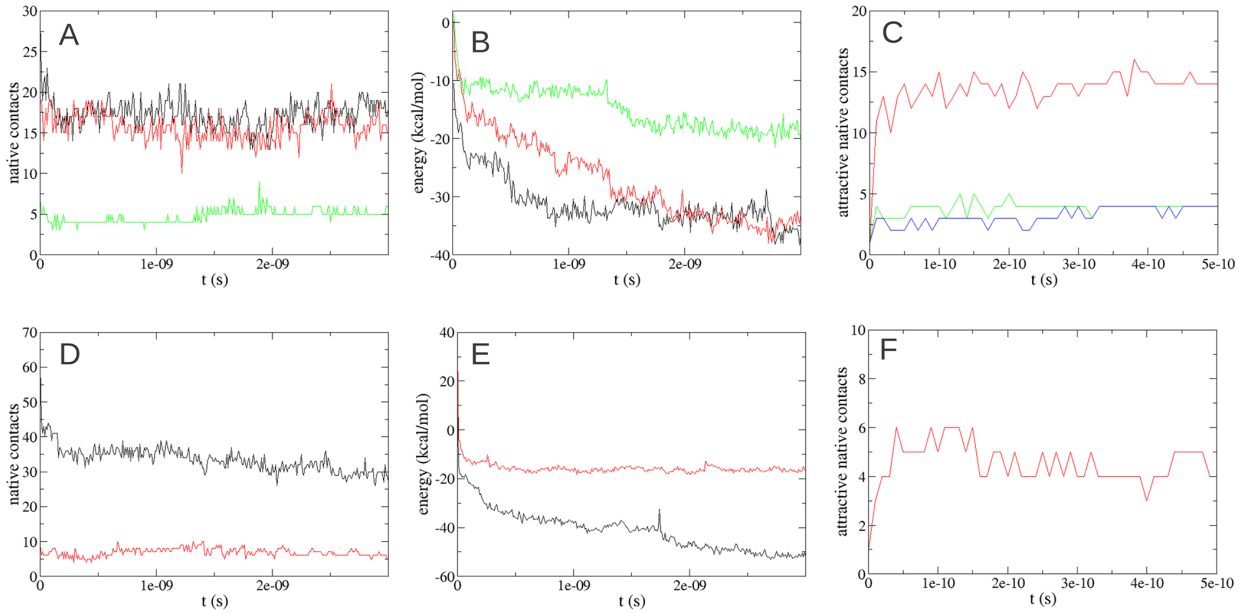


Figure 4. Analysis of DMD relaxation trajectories for two successful cases. (A) Native contacts, (B) energy, and (C) native attractive contacts along DMD relaxation for different conformations in the 1Z0K case. (D, E, F) Same for 1M10 case. The color code is as follows: experimental complex structure in black; best near-native conformation after DMD relaxation in red; other near-native conformations in green; other non-near-native conformations in blue.

the ones that most benefit from DMD relaxation. However, proteins that do not change conformation upon binding do not need DMD relaxation, as rigid-body docking already gives good predictions for them. Thus, in a realistic situation it would be interesting to know whether DMD relaxation is going to be useful for a given case or not. For that, we can assume that flexibility of proteins is related to the conformational variability upon binding; hence we would just need a method to estimate conformational flexibility of the individual proteins. Here, we have estimated the extent of the conformational motion of a protein based on NMA[16] (see Outline of the Method). It is important to notice that this method to estimate the extent of the conformational changes is valid as long as the normal mode hypothesis is valid for a given protein; therefore it might happen that a protein has a conformational change upon binding that is very different from the change predicted from NMA. In addition, global flexibility of a protein based on NMA might not need to correlate with interface deformation upon binding, which has been shown here to be critical for the performance of DMD. Figure 5 shows the top 10 success rates after DMD relaxation for the cases that were classified as rigid or flexible based on normal mode calculations. Interestingly, the improvement of success rates after DMD relaxation for the predicted flexible cases is evident, virtually the same that we obtained for the real flexible cases. This result indicates that, based on simple normal mode calculations, one could estimate whether DMD relaxation is going to be useful for a given protein−protein docking problem.

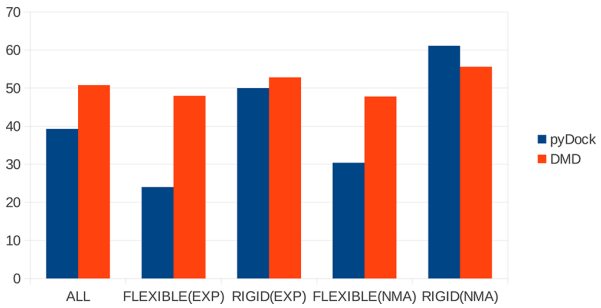


Figure 5. Top 10 success rates for the predicted rigid and flexible cases according to normal-mode analysis. For comparison, the top 10 success rates for the known rigid and flexible cases, as well as for the whole data set, are also shown.

CONCLUSIONS

We outlined here a fast, simple, and efficient protocol for protein−protein docking, which combines rigid-body orientation sampling with structural relaxation by means of discrete molecular dynamics (DMD) simulations. The procedure takes advantage of the intrinsic characteristics of DMD, such as (i) the use of nondifferentiable square potentials, typically used in docking procedures but difficult to implement in Newtonian molecular dynamics, (ii) the high computational efficiency of the algorithm to trace large scale movements, (iii) the simplicity of using the method in the context of multiresolution representations of the proteins, and (iv) the ability of the method to maintain intramolecular geometry in some parts of the monomer, while keeping intact flexibility in other regions and maintaining the global inter-residual flexibility. We have found that the predictive power of the DMD relaxation method is much less affected by the deformation upon binding of the ligand and receptor as compared to rigid-body docking methods, leading to a clear improvement in the predictive success rates over the complete benchmark. Finally, we found that it is possible to estimate a priori whether a given case is going to benefit from DMD relaxation.

AUTHOR INFORMATION

Corresponding Author

*E-mail: modesto.orozco@irbbarcelona.org (M.O.) or juanf@bsc.es (J.F.-R.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Laura Orellana for help with the NMA calculations. This work has been supported by grant number BIO2010-22324 (J.F.-R.) and BIO2009-10964 (M.O.) from MICINN-Spain, the European Research Council (ERC-Advanced Grant, M.O.), the Instituto Nacional de Bioinformatica (INB; M.O., J.L.G.), the Consolider E-Science Project (M.O.), and the Fundación Marcelino Botin (M.O.). P.S. is a fellow of the La Caixa doctoral program. M.O. is an ICREA Academia Fellow.

REFERENCES

(1) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535−542.

(2) Stein, A.; Mosca, R.; Aloy, P. *Curr. Opin. Struct. Biol.* **2011**, *21*, 200−208.

(3) Venkatesan, K.; Rual, J.-F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K.; Yildirim, M. A.; Simonis, N.; Heinzmann, K.; Gebreab, F.; Sahalie, J. M.; Cevik, S.; Simon, C.; Smet, A.-S.; Dann, E.; Smolyar, A.; Vinayagam, A.; Yu, H.; Szeto, D.; Borick, H.; Dricot, A.; Klitgord, N.; Murray, R. R.; Lin, C.; Lalowski, M.; Timm, J.; Rau, K.; Boone, C.; Braun, P.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Tavernier, J.; Wanker, E. W.; Barabási, A.-L.; Vidal, M. *Nat. Methods* **2009**, *6*, 83−90.

(4) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins* **2002**, *47*, 409−443.

(5) Ritchie, D. *Curr. Protein Pept. Sci.* **2008**, *9*, 1−15.

(6) Pons, C.; Grosdidier, S.; Solernou, A.; Pérez-Cano, L.; Fernández-Recio, J. *Proteins* **2010**, *78*, 95−108.

(7) Fischer, H. E. *Chem. Ber.* **1894**, *27*, 2985−2993.

(8) Stein., A.; Rueda, M.; Panjkovich, A.; Orozco, M.; Aloy, P. *Structure* **2011**, *19*, 881−889.

(9) Pons, C.; D'Abramo, M.; Svergun, D. I.; Orozco, M.; Bernadó, P.; Fernández-Recio, J. *J. Mol. Biol.* **2010**, *403*, 217−230.

(10) Zacharias, M. *Curr. Opin. Struct. Biol.* **2010**, *20*, 180−186.

(11) Bonvin, A. M. J. J. *Curr. Opin. Struct. Biol.* **2006**, *16*, 194−200.

(12) Fernández-Recio, J.; Totrov, M.; Abagyan, R. *Protein Sci.* **2002**, *11*, 280−291.

(13) Fernández-Recio, J.; Totrov, M.; Abagyan, R. *Proteins* **2003**, *52*, 113−117.

(14) Fernández-Recio, J.; Abagyan, R.; Totrov, M. *Proteins* **2005**, *60*, 308−313.

(15) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. *J. Am. Chem. Soc.* **2003**, *125*, 1731−1737.

(16) Dobbins, S. E.; Lesk, V. I.; Sternberg, M. J. E. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 10390−10395.

(17) Zacharias, M. *Proteins* **2004**, *54*, 759−767.

(18) Wang, C.; Bradley, P.; Baker, D. *J. Mol. Biol.* **2007**, *373*, 503−519.

(19) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. *J. Mol. Biol.* **2003**, *331*, 281−299.

(20) Chen, R.; Mintseris, J.; Janin, J.; Weng, Z. *Proteins* **2003**, *52*, 88−91.

(21) Zhou, Y. Q.; Karplus, M. *Nature* **1999**, *401*, 400−403.

(22) Emperador, A.; Meyer, T.; Orozco, M. *Proteins* **2010**, *78*, 83−94.

(23) Urbanc, B.; Cruz, L.; Yun, S.; Buldyrev, S. V.; Bitan, G.; Teplow, D. B.; Stanley, H. E. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17345−17350.

(24) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147−155.

(25) Ding, F.; Sharma, S.; Chalasani, P.; Demidov, V. V.; Broude, N. E.; Dokholyan, N. V. *RNA* **2008**, *14*, 1164−1173.

(26) Ding, F.; LaRoque, J. J.; Dokholyan, N. V. *J. Biol. Chem.* **2005**, *280*, 40235−40240.

(27) Nguyen, H.; Hall, C. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 16180−16185.

(28) Ding, F.; Borreguero, J. M.; Buldyrev, S. V.; Stanley, H. E.; Dokholyan, N. V. *Proteins* **2003**, *53*, 220−228.

(29) Sfriso, P.; Emperador, A.; Orellana, L.; Hospital, A.; Gelpi, J. L.; Orozco, M. *J. Chem. Theory Comput.* **2012**, *8*, 4707−4718.

(30) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459−466.

(31) Smith, W. S.; Hall, C. K.; Freeman, B. D. *J. Comput. Phys.* **1997**, *134*, 16−30.

(32) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. *Proteins* **2010**, *78*, 3111−3114.

(33) Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. E. *J. Mol. Biol.* **1997**, *272*, 106−120.

(34) Cheng, T.; Blundell, T. L.; Fernández-Recio, J. *Proteins* **2007**, *68*, 503−515.

(35) Grosdidier, S.; Pons, C.; Solernou, A.; Fernández-Recio, J. *Proteins* **2007**, *69*, 852−858.

(36) Pons, C.; Solernou, A.; Pérez-Cano, L.; Grosdidier, S.; Fernández-Recio, J. *Proteins* **2010**, *78*, 3182−3188.

(37) Fernández-Recio, J.; Totrov, M.; Abagyan, R. *J. Mol. Biol.* **2004**, *335*, 843−865.

(38) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133−152.

(39) Rueda, M.; Chachon, P.; Orozco, M. *Structure* **2007**, *15*, 565−575.

(40) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Pérez, A.; Camps, J.; Hospital, A.; Gelpí, J. L.; Orozco, M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 796−801.

7. 3 PACSAB: Coarse-Grained Force Field for the Study of Protein–Protein Interactions and Conformational Sampling in Multiprotein Systems

Context:

Atomistic protein simulations are usually performed on single molecules for efficiency purpose. Coarse-grained models that could handle larger systems are often calibrated against single-molecule simulations, giving a poor description of intermolecular interactions. We present a new coarse-grained force field for the study of multi-protein systems. The force field, which is implemented in the context of the dMD algorithm, is able to reproduce the properties of folded and unfolded proteins, in either isolation or forming well-defined complexes. Our approach also reproduces aggregated proteins thanks to its proper evaluation of protein–protein interactions. The accuracy and computational efficiency of the method makes it an useful tool to study molecular processes involving many proteins, with particular focus on aggregation-based diseases.

Title: PACSAB: Coarse-Grained Force Field for the Study of Protein–Protein Interactions and Conformational Sampling in Multiprotein

Authors: Agustí Emperador, Pedro Sfriso, Marcos Villareal, Josep Lluis Gelpi and Modesto Orozco

Stage: Published

Journal: Journal of Chemical Theory and Computation

Type: Research Article

Supplementary Material: http://pubs.acs.org/doi/abs/10.1021/acs.jctc.5b00660

Author Contribution: P.S contributed to design research and analysed results.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

PACSAB: Coarse-Grained Force Field for the Study of Protein−Protein Interactions and Conformational Sampling in Multiprotein Systems

Agustí Emperador,*[,†,‡] Pedro Sfriso,[†,‡] Marcos Ariel Villarreal,[§] Josep Lluis Gelpí,[†,‡,∥,⊥] and Modesto Orozco*[,†,‡,∥,⊥]

[†]Institute for Research in Biomedicine (IRB Barcelona), Baldiri i Reixac 10, Barcelona 08028, Spain

[‡]Joint BSC-IRB Research Program in Computational Biology, IRB Barcelona, Barcelona 08028, Spain

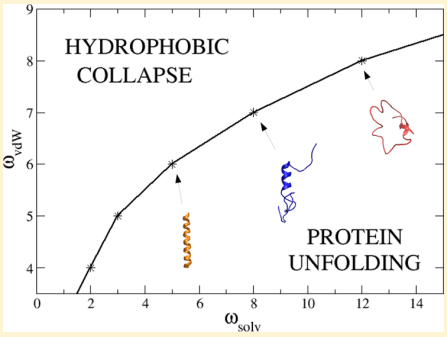
[§]Instituto de Investigaciones en Fisicoquímica de Córdoba - Departamento de Matemática y Física, CONICET-Universidad Nacional de Córdoba, University City, Córdoba 5000, Argentina

[∥]Barcelona Supercomputing Center, Jordi Girona 29, Barcelona 08034, Spain

[⊥]Departament de Bioquímica, Facultat de Biologia, Avgda Diagonal 645, Barcelona 08028, Spain

Supporting Information

ABSTRACT: Molecular dynamics simulations of proteins are usually performed on a single molecule, and coarse-grained protein models are calibrated using single-molecule simulations, therefore ignoring intermolecular interactions. We present here a new coarse-grained force field for the study of many protein systems. The force field, which is implemented in the context of the discrete molecular dynamics algorithm, is able to reproduce the properties of folded and unfolded proteins, in both isolation, complexed forming well-defined quaternary structures, or aggregated, thanks to its proper evaluation of protein−protein interactions. The accuracy and computational efficiency of the method makes it a universal tool for the study of the structure, dynamics, and association/dissociation of proteins.



I. INTRODUCTION

The theoretical representation of systems of interacting proteins presents major challenges due to the need to simulate very large systems (often above millions of atoms) for very long periods of time (in some cases on the time scale of days[1]). Despite the impressive advance of atomistic molecular dynamics,[2] the representation of protein structure, dynamics, and interactions still needs of the use of simplified models that allow a more efficient sampling of the protein conformational space. Coarse-grained (CG) models increase computational efficiency by using implicit solvent models[3] and by collapsing groups of atoms on beads.[4] This results in a reduction of the number of degrees of freedom of the system, which combined with more efficient motion propagation schemes accelerates the simulations with respect to atomistic molecular dynamics.

Most transferable CG force fields for proteins were fitted to reproduce the folded state of a protein,[5−7] or at most to reproduce the transition from unfolded to folded state,[8−12] but they cannot reproduce the behavior of intrinsically disordered proteins (IDPs). Attempts to develop IDP CG models yield to functionals which are unable to represent folded proteins,[13,14] highlighting the problems to represent with a single functional folded and unfolded states of proteins. Furthermore, existing CG force fields were created to study isolated proteins and are not prepared to reproduce well-ordered protein complexes. Some of the most successful coarse-grained models used in molecular dynamics simulations of proteins have been PaLaCe[5] and PRIMO,[7] that give an excellent description of the structure and dynamics of folded proteins, and also OPEP[11] which, apart from that, was able to fold several peptides and sample conformational changes in small aggregates.[12] In summary, despite decades of effort, there are not general CG methods able to represent correctly the dynamics of proteins both in its folded and unfolded conformations, and the association/dissociation dynamics in multiprotein systems. This lack of methodology hampers our ability to describe theoretically the dynamics, interactions, and association of proteins.

We present here a **p**airwise **a**dditive potential for **c**oarse-grained **s**ide chains and **a**tomistic **b**ackbone protein model (PACSAB) with a transferable force field for the simulation of many-protein systems. Contrary to many CG models which are based on knowledge rules on folded proteins,[15] our approach is based on a contraction of an implicit solvent classical atomistic model, which makes possible transferability to different sce-

Received: July 10, 2015
Published: November 10, 2015

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

narios and systems. The force field is adapted to the framework of discrete molecular dynamics (DMD),[16] which allows a very efficient sampling of large protein systems. The parameters defining the potential energy function in the model were fitted by exploring the phase diagram for a solution of Aβ40 peptides. The resulting force field was then tested in DMD simulations of IDPs, folded proteins, and protein−protein complexes. In summary, we present here a universal coarse-grained simulation model to explore the conformational space and interactions in multiprotein systems.

II. METHODS

Mapping of the Proteins. The aim of our model is to study different conformations and aggregation states of proteins, which means that the coarse-graining strategy should be designed to reproduce accurately excluded volume effects, side chain packing and backbone hydrogen bonding. Following Marrink's strategy,[17,18] we have placed beads at all C_αs to define the protein trace, plus a variable number of beads to describe the side chains using the mapping defined in ref 19 (typically each bead represents four heavy atoms; see Figure 1A).

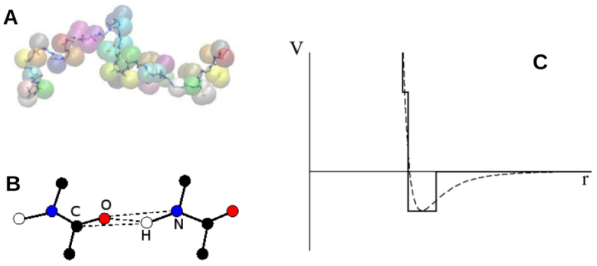


Figure 1. (A) Extended conformation of a Aβ40 peptide in our coarse-grained model. Each residue is represented with a different color. (B) Pseudobonds used to define the hydrogen bond (see main text). (C) Schematic picture of the construction of the discretized potential (solid line). The potential well is centered around $R^*_{AB} = R^*_A + R^*_B$, the sum of the bead radii (see main text). The dashed line is the continuous potential.

We have concentrated all the atoms of the bead on its center of mass, therefore all the atom−atom distances become equal to the bead−bead distances. Additionally, to represent explicitly hydrogen bonds we have added the backbone atoms N, H, C, and O.[5] We have also added a dummy atom bound to the C_α of each residue in order to keep the proper chirality of the amino acids. Solvent effects were reproduced using an implicit solvent model, which increases computational efficiency and sampling in the study of diluted systems.

DMD Simulations and Sampling. In DMD simulations the particles are considered as hard spheres interacting through discontinuous potentials, therefore moving at constant velocity until a collision (event) happens.[16] Events occur when pairwise distance equals the distance of a discontinuity in the interaction potential (see Figure 1C). No forces have to be calculated, and it is not necessary to integrate the equations of motion, speeding up the simulation as compared with conventional molecular dynamics (MD). Hardcore potentials preventing steric clashes are defined between unbound particles, and infinite square wells are defined between bound particles to keep the proper bond distances. Additional square wells are used to preserve the side chain geometry (pseudobonds).

According to DMD, the trajectory of the particles between collisions is

$$\vec{r}_i(t + t_c) = \vec{r}_i(t) + \vec{v}_i(t)t_c$$

where $t_c = \min(t_{ij})$ is the next collision time and

$$t_{ij} = \frac{-\vec{r}_{ij}\cdot\vec{v}_{ij} \pm \sqrt{(\vec{r}_{ij}\cdot\vec{v}_{ij})^2 - v_{ij}^2(r_{ij}^2 - d^2)}}{v_{ij}^2}$$

being d the sum of the radii of particles i and j.

When two particles collide there is a transfer of linear momentum

$$m_i\vec{v}_i = m_i\vec{v}_i' + \Delta\vec{p}$$

$$m_j\vec{v}_j + \Delta\vec{p} = m_j\vec{v}_j'$$

Conservation of momentum and energy is imposed at each event, and from this, the velocity of each particle after the collision is found

$$m_i\vec{v}_i + m_j\vec{v}_j = m_i\vec{v}_i' + m_j\vec{v}_j'$$

$$\frac{1}{2}m_i v_i^2 + \frac{1}{2}m_j v_j^2 = \frac{1}{2}m_i v_i'^2 + \frac{1}{2}m_j v_j'^2 + \Delta V$$

Simulations were performed in the canonical ensemble, using the Andersen thermostat (for more details, see ref 16). The sampling obtained in implicit solvent DMD simulations is much higher than that expected from atomistic explicit solvent MD, due to the lack of collisions with solvent molecules. In practice, this implies that simulation time defined in a DMD trajectory corresponds to roughly 2−3 orders of magnitude longer real time.[20] The speed of the DMD simulations with the PACSAB model for different systems studied in this work is shown in Table S1 of the Supporting Information.

Construction of the Force Field. The force field consists of bonded and nonbonded terms. The first set of terms is used to maintain covalent structure, while the second accounts for intra- and intermolecular interactions. In all the cases, the different terms of the interaction potential are expressed by means of square well potentials to make possible their implementation with the DMD algorithm.

Bonded Terms. Square potentials are used to maintain all chemical bonds and to fix the bond angles. We also use a pseudobond to fix the dihedral angle of the peptide bonds in order to enforce its planar geometry, but we do not implement any other dihedral in the PACSAB model. Bonds and pseudobonds are defined as narrow square wells (with infinite depth to prevent bond breaking), whose center is at the equilibrium distance corresponding to each covalent bond, angle or dihedral.[21]

Nonbonded Terms. The non-bonded interactions comprise hydrogen bonding, defined only between the amide N, H, C, and O atoms and interactions between nonbonded beads (van der Waals and implicit solvation), affecting only C_α and side chain beads.

Hydrogen Bonds. They are represented by means of square wells of depth E_{hb} and are defined for the pairs O−H, O−N, and C−H, whenever these four atoms fulfill a geometry corresponding to the correct alignment and distance between the two dipoles N−H and O−C (see Figure 1B). Following the ideas in ref 22, we increase the stability of hydrogen bonds that are buried inside the protein, therefore not distorted by interactions with water. With this purpose the hydrogen bond

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

energy is defined as $E_{hb} = E_{hb}\gamma_i\gamma_j + E_{hb}^{core}(1 - \gamma_i\gamma_j)$, where $E_{hb}^{core} > E_{hb}$. γ is a structurally dependent shifting function that helps to smoothly move from fully exposed ($\gamma = 1$) to buried ($\gamma = 0$):

$$\gamma(n) = \frac{1}{1 + \exp((n - \alpha)/\beta)} \quad (1)$$

where n is an integer quantity related to the level of exposition to the solvent of the particle (see Appendix A), α is the limit value between exposed and buried, and β is the sharpness of the step. The values of α and β were adjusted from simulations on our training set of folded proteins (see below).

Interactions between Nonbonded Beads. The interaction between any pair of nonbonded beads A and B is defined as

$$V_{AB}(r) = \omega_{vdW}V_{AB}^{vdW}(r) + \omega_{solv}V_{AB}^{solv}(r) \quad (2)$$

where ω_{vdW} and ω_{solv} are the weights of the optimized van der Waals and implicit solvation terms (see below). In this work we have considered that electrostatic effects are properly included[22] through the hydrogen bonding and the implicit solvation terms Such an approach was used with success in the ab initio folding of several small proteins.[22]

To construct the interactions between non bonded coarse-grained beads, we assume that all the nonbonded interaction potential terms are pairwise additive in terms of the atomistic interactions:

$$V_{AB}(r) = \sum_{i\in A}\sum_{j\in B} V_{ij}(r) \quad (3)$$

where r is the distance between beads A and B and $V_{ij}(r)$ is the atomistic interaction. We use the van der Waals parameters ε_i^* of the atomistic CHARMM19 force field[23] to construct the coarse-grained van der Waals interactions, plus the atomistic EEF1 effective energy function of Lazaridis and Karplus[24] to derive the implicit solvent coarse-grained model. Constructing our potentials from atomistic interactions allows us to avoid biases[25] due to the use of statistical potentials derived from databases of crystal structures,[26,27] opening the possibility to study disordered proteins.

The *van der Waals interaction* between the coarse-grained beads is defined as

$$V_{AB}^{vdW}(r) = \varepsilon_{AB}^*\left[\left(\frac{R_{AB}^*}{r}\right)^{12} - 2\left(\frac{R_{AB}^*}{r}\right)^{6}\right] \quad (4)$$

r being the distance between beads A and B. $R_{AB}^* = R_A^* + R_B^*$, the sum of the radii of beads A and B. To compute the bead radii, we consider that the volume of each bead is proportional to the sum of the volumes of each atom included into the bead, leading to the relation $R^* = \rho(\sum R_i^{*3})^{1/3}$, R_i^* being the radius of each atom, ρ being fitted to 0.9 after inspection of atomistic residue−residue interaction profiles.

The interaction hardness ε_{AB}^* is computed extrapolating from atomistic van der Waals interactions (see Appendix B):

$$\varepsilon_{AB}^* = -\sqrt{\sum_{i\in A}\varepsilon_i^* \sum_{j\in B}\varepsilon_j^*}\left[\left(\frac{2/\rho}{N_A^{1/3} + N_B^{1/3}}\right)^{12} - 2\left(\frac{2/\rho}{N_A^{1/3} + N_B^{1/3}}\right)^{6}\right] \quad (5)$$

where N_A (N_B) is the number of atoms included by bead A (B) and ε_i^* are the atomistic van der Waals interaction hardnesses.

The *implicit solvation term* between the coarse-grained beads has been constructed from the atomistic EEF1 functional[24]

$$V_{ij}^{solv}(r) = -\int_{v_i} f_j\, d\vec{r} - \int_{v_j} f_i\, d\vec{r} \approx -f_j(r)v_i - f_i(r)v_j \quad (6)$$

where v_i is the volume and $f_i(r) = C\Delta G_i \exp(-(r/\lambda)^2)/r^2$ is the solvation free energy density of particle i, ΔG_i being the solvation free energy of the isolated atom i, λ a correlation length and $C = 1/(2\pi^{3/2}\lambda)$.[24] Both ΔG_i and v_i for each particle type are determined from experimental data.[24] The previous equation can be rewritten as

$$V_{ij}^{solv}(r) \approx -C(\Delta G_i v_j + \Delta G_j v_i)\exp(-(r/\lambda)^2)/r^2 \quad (7)$$

Assuming additivity, the solvation term affecting beads A and B is then defined as

$$V_{AB}^{solv}(r) \approx -C\sum_{i\in A}\sum_{j\in B}(\Delta G_i v_j + \Delta G_j v_i)\exp(-(r/\lambda)^2)/r^2 \quad (8)$$

We have included more information about the parameters of the atomistic force fields CHARMM19 and EEF1 in the Supporting Information.

The EEF1 implicit solvation functional assumes[24] that any "nonprotein space" is "solvent space", even if it is inside the protein. Thus, this model does not take into account that water has a finite size and cannot fit inside the core of the protein. This can be quite realistic when using an atomistic representation of the protein, but this is not a good approximation for coarse-grained representations of the system, where packing in the interior of the protein cannot be as dense. To correct this spureous effect we have modulated the implicit solvation term by including the factor γ (eq 1) in eq 8:

$$V_{AB}^{solv}(r) \approx -C(\gamma_A\sum_{i\in A}\Delta G_i\sum_{j\in B}v_j + \gamma_B\sum_{j\in B}\Delta G_j\sum_{i\in A}v_i)\exp(-(r/\lambda)^2)/r^2 \quad (9)$$

$$V_{AB}^{solv}(r) \approx -C(\Delta G_A v_B + \Delta G_B v_A)\exp(-(r/\lambda)^2)/r^2 \quad (10)$$

where $\Delta G_A = \gamma_A\sum_{i\in A}\Delta G_i$ and $v_A = \sum_{i\in A} v_i$.

Discretization of the Total Interaction Potential between Nonbonded Coarse-Grained Beads. To transform the potential described above to a discretized functional which can be inserted in the DMD algorithm, we define a well located at $R_{AB}^* = R_A^* + R_B^*$ (the minimum of the coarse-grained van der Waals potential term; see Figure 1). The well depth is computed as the sum of the two terms at distance $r = R_{AB}^*$:

$$V_{AB}(R_{AB}^*) = \omega_{vdW}V_{AB}^{vdW}(R_{AB}^*) + \omega_{solv}V_{AB}^{solv}(R_{AB}^*) \quad (11)$$

To reduce the computational cost of the simulations we approximate the nonbonded potential of mean force to a discretized potential with two energy steps, that form a potential well (or barrier) if the total potential is attractive (or repulsive). The inner and outer step distances are $0.9R_{AB}^*$ and $1.1R_{AB}^*$, respectively, while the hardcore repulsion distance was placed at $0.88R_{AB}^*$ (see Figure 1C).

Parametrization of the Force Field. We refined the parameters of the force field by analyzing the behavior in water of a single Aβ40 peptide, a 30 μM solution of Aβ40 peptides,[28] and a small folded protein (PDB id 1FAS). Our objective was

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

to find a parametrization able to represent correctly the three states (unfolded, aggregated, and folded).

We used the simulations of the protein 1FAS to adjust the hydrogen bonding strengths E_{hb} = 3 kcal/mol and E_{hb}^{core} = 4 kcal/mol as well as the parameters α = 10 and β = 0.5 used to define the factor γ (see eq 1).

The values of ω_{solv} and ω_{vdW} in eq 11 were selected to get a proper balance between aggregation and dissociation rates in simulations of a solution of Aβ40 peptides at a concentration of 30 μM.

III. RESULTS AND DISCUSSION

Force Field Calibration. The macroscopic solution was modeled by placing four Aβ40 peptides in a cubic box of the size corresponding to 30 μM concentration and with periodic boundary conditions. We observed that in this system the solvation term prompts dissociation and the van der Waals term prompts association. In order to have a good statistics of the association process we ran eight long DMD simulations for each point in the two-dimensional space (ω_{solv}, ω_{vdW}). We scanned the range of ω_{vdW} from 0 to 10 and ω_{solv} from 0 to 18 (mesh density of one unit per dimension) to build the phase diagram shown in Figure 2. Above the phase boundary line,

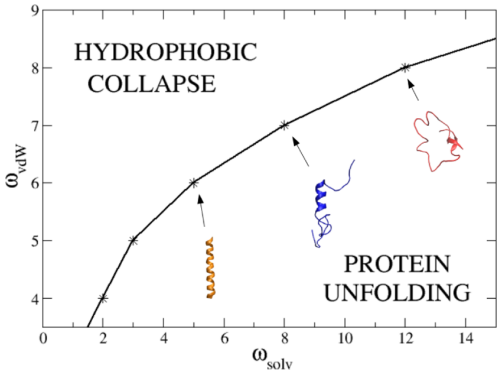


Figure 2. Phase diagram for the 30 μM concentration Aβ40 solution. Above the phase boundary line the solution precipitates. Small pictures show typical secondary structures obtained for monomeric Aβ40 in the simulations at certain points on the phase boundary line.

aggregation happens due to hydrophobic collapse, and below it, there is equilibrium between peptide associations and dissociations. This stationary regime was achieved when the trajectories reached 3 μs, but to make sure that the oligomer size distribution was stabilized we extended the simulations up to 5 μs. Equilibrium is rapidly reached in the DMD simulations due to the enhanced conformational sampling of the implicit solvent model we use, making 1 μs equivalent to 1 ms of real time (see Methods).

We chose the point ω_{solv} = 8, ω_{vdW} = 7 on the phase boundary line, that gives the secondary structure in better agreement with the conformational sampling obtained in the explicit solvent atomistic MD simulation of a single Aβ40 peptide in ref 29 (see structures in Figure 2), as well as a realistic aggregation profile in the Aβ40 solution. We refer the reader to ref 30 for a recent review about simulations of the Aβ40 peptide. We show in Figure S1 of the Supporting Information the secondary structure evolution as a function of time for an Aβ40 peptide. We started the simulation from a completely extended conformation, and rapidly an α-helix region is formed between residues 12 and 24. We show the α-helix and β-strand propensities in Figure S2. These results are in agreement with those obtained from united-atom implicit solvent simulations in refs 31 and 32 that give a higher stability of these secondary structure regions as compared with the explicit solvent simulations of ref 33.

In order to test the stability of α-helix and β-sheet motifs with the PACSAB force field, we have made simulations of an α-helix peptide (EK peptide) and of a β-sheet peptide (the Gly5-Trp29 segment of the protein with PDB code 1I6C), both starting from completely extended conformations. PACSAB folded these peptides to their native conformation, as shown in the Supporting Information (Figure S5).

We have shown in Figure 3 the evolution with time of the population of each oligomer order (computed as the average

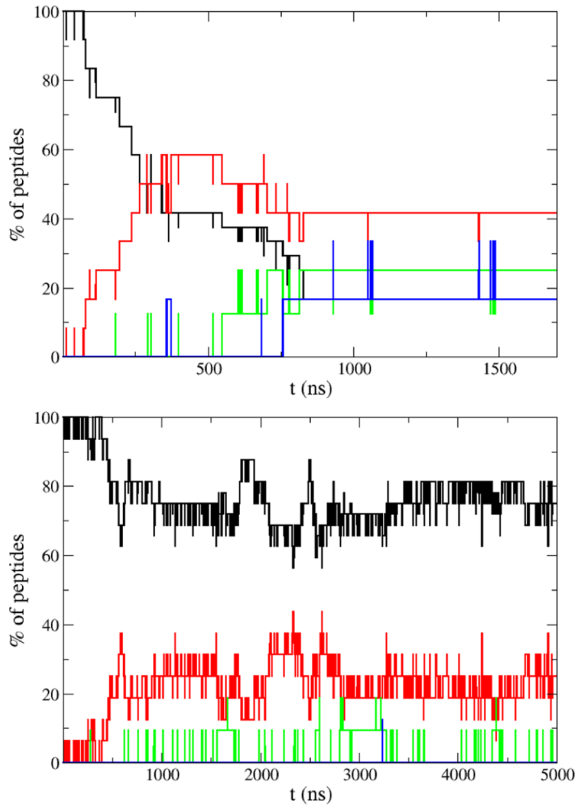


Figure 3. Evolution of the percentage of peptides in each oligomeric state during the trajectory (black line monomers; red line dimers; green line trimers; blue line tetramers): (upper panel) evolution for the point at coordinates (6,7) in the phase diagram; (lower panel) same for the point at (8,7).

over the eight simulations) for this point and for a point above the phase boundary line, where the dynamics of aggregation tends to populate higher order oligomers. We selected the protein 1FAS as a training system for fine-tuning of the (ω_{solv}, ω_{vdW}) values. However, as can be observed in Figure 4, no reoptimization was necessary, since the chosen parametization reproduces correctly the structure of this folded protein. If a (ω_{solv}, ω_{vdW}) value below the phase boundary line in Figure 2 is chosen, the protein unfolds due to the underestimated hydrophobicity with such parametrization.

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

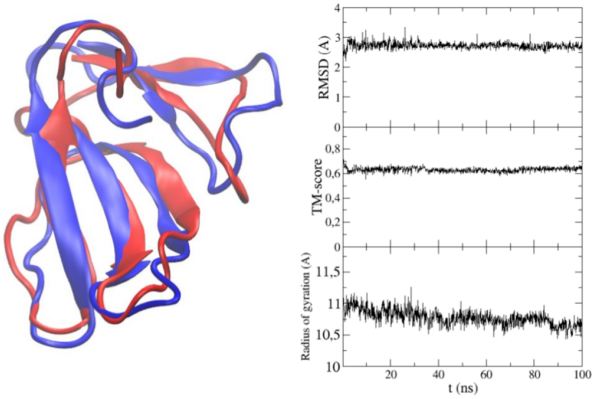


Figure 4. Structure of the protein with PDB id 1FAS after a DMD simulation of 100 ns (red cartoon), compared with the crystallographic structure (blue). Also shown are the RMSD, TM-score, and radius of gyration.

Amyloid Aggregation Dynamics. For each point in the phase diagram of Figure 2 we started the simulations from completely extended conformations of the peptides. We observed that at the beginning of the simulation, the peptides experience a fast collapse that drive their structure to a fold intermediate between a helical structure and a molten globule (see Figure 5), in good agreement with previous explicit solvent atomistic MD simulations of Aβ40.[29]

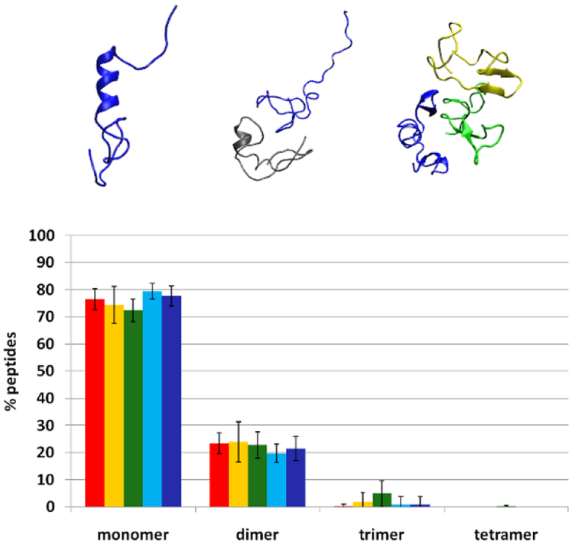


Figure 5. Oligomerization in the 30 μM Aβ40 peptide solution. (upper panel) Structures of a monomer, a dimer, and a trimer obtained during the simulations. (lower panel) Percentage of peptides in each oligomeric state observed at different DMD simulation times using the optimal force field parametrization (see main text): 1 μs (red), 2 μs (yellow), 3 μs (green), 4 μs (blue), and 5 μs (dark blue).

As simulation progresses, intermolecular collisions happen, some of them leading to peptide association. At 30 μM concentration peptides collide every 0.1 μs on average, but only ~10% of these collisions are productive (leading to the formation of a stable dimer). The low frequency of association/dissociation events requires an extensive sampling than cannot be achieved by standard explicit solvent atomistic MD simulations, but that was accessible with our implicit solvent DMD simulations (note in Figure 5 that a stationary regime had been reached within the simulation window). The association of monomers with dimers led to the formation of trimers, much less abundant due to the low population of dimers. For the same reason the existence of tetramers was residual. Our oligomer size distribution is coincident with the experimental distributions observed in a very recent work by Pujol-Pina et al.[34] (see Figures S3 and S4 of the Supporting Information)

We made simulations at higher concentrations, finding that higher order oligomers become more abundant as the concentration increases. Figure 6 shows the oligomer size distribution

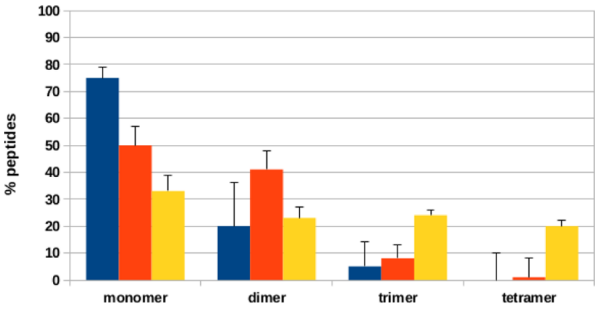


Figure 6. Percentage of peptides in each oligomeric state at different concentrations: 50 μM (blue), 100 μM (orange), and 240 μM (yellow).

obtained after 1 μs DMD simulations at different concentrations. Eight simulations were performed for each concentration. The oligomer size distribution at 50 μM fits well with the distribution at 30 μM (see Figure 5), but at 100 μM it has changed clearly with an evident increase of dimers. At 240 μM similar populations are found for monomers, dimers, trimers, and tetramers. This tendency is consistent with the results of very recent atomistic molecular dynamics simulations for solutions of β-amyloid peptides at very high concentration.[35]

Test Systems. In order to evaluate the quality and universality of the force field, we performed a comprehensive evaluation for folded proteins, intrinsically disordered proteins and protein−protein complexes.

Folded Proteins. We explored the ability of the coarse grained force field to reproduce the structure of folded proteins in long simulation time scales. For this purpose we selected a set of 25 proteins representative of the most prevalent protein folds[36] and performed DMD simulations of 500 ns (this gives, as explained above, a sampling corresponding to multimicrosecond trajectories in explicit solvent atomistic MD). Results in Figure 7 show that all the trajectories are stable, without any evident signal of unfolding as illustrated in the evolution of the radii of gyration. The RMS deviations from experimental structure are typically in the range 2−8 Å, higher than those found in atomistic MD simulations,[36−38] but matching the level of accuracy of state-of-the-art CG force fields designed specifically to reproduce the folded state of proteins[5,7,12] (see comparison with other coarse-grained models in Table S2 of the Supporting Information)

As demonstrated by the TM-score value,[36] the flexible loops are the main origin of the deviation of DMD samplings from experimental structures, while the protein core is fully

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

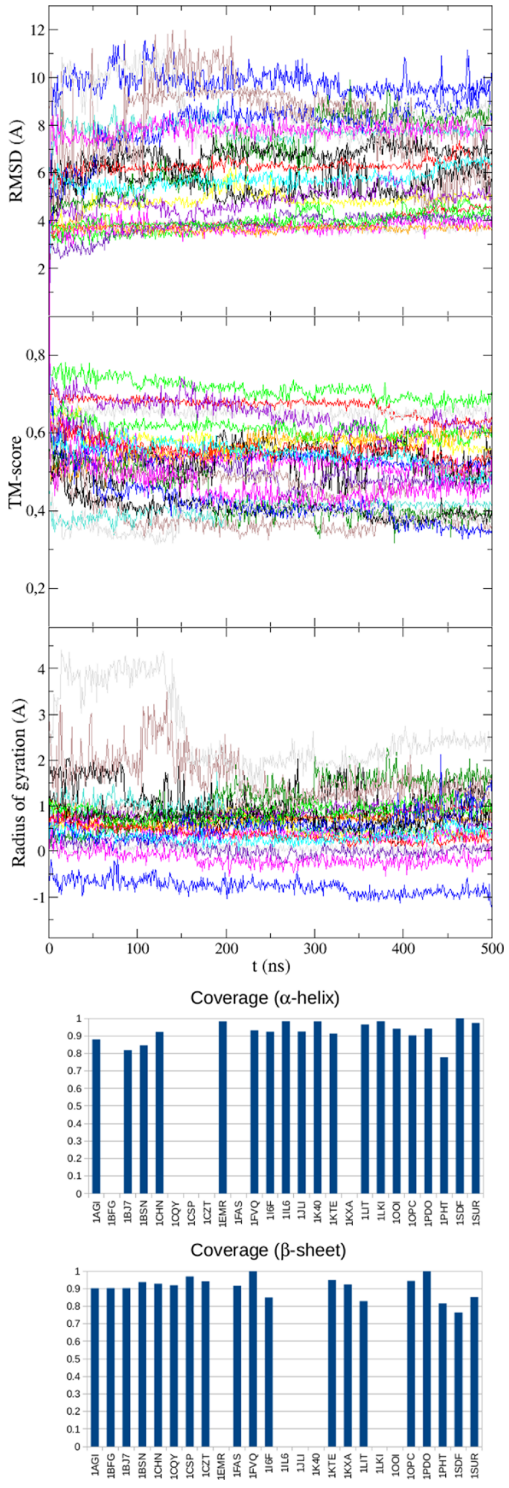


Figure 7. Structural properties of the folded protein benchmark after a DMD simulation of 500 ns. (top to bottom) RMSD respect to the native conformation, TM-score, change in the radius of gyration during the trajectory, conserved α-helix, and conserved β-sheet. The value of RMSD, TM-score, and radius of gyration is plotted in a different color for each protein. Proteins without α-helices are not shown in the α-helix coverage plot, and the same for β-sheets.

preserved during the simulations. We found conservation of native secondary structure (Figure 7), which is remarkable considering that our force field does not introduce, like others, specific restraints or backbone dihedral terms favoring the stability of usual secondary structure elements. Despite this lack of restrictions in the backbone dihedrals, the Ramachandran plots are well reproduced, as shown in Figure S6 of the Supporting Information. Finally, fold recognition algorithms detected in all the cases the real structure (or that of a very close homologue) from the sampled structure at the end of the DMD simulation (see Figures S7 and S8 of the Supporting Information)

In summary, despite the lack of specific training for folded proteins, the absence of restrictions on secondary structure, and the lack of structure potentials biasing the simulations toward the native state, our extremely simple CG model is able to sample properly the structure of folded proteins in very long simulations.

Intrinsically Disordered Proteins. To test the generality and universality of the force field we also explored the dynamics of two intrinsically disordered proteins (IDPs): ACTR and α-synuclein. ACTR is an IDP that folds in a well-defined structure only in the presence of its macromolecular partner,[39] while in its absence appears as a random coil with a residual percentage of α-helix. DMD simulations recognized the IDP nature of ACTR, sampling a wide variety of conformational states in a 1 μs DMD trajectory (see Figure 8). The only common feature between the conformations of the ensemble is the formation of residual secondary structure, in good agreement with circular dichroism measurements.[40]

Similar success is obtained with α-synuclein, which is stable when embedded in a lipid environment, while it is disordered in water,[41] except for some residual contacts between the residues around position 50 and the residues around position 120 in the sequence. As shown in Figure 8, the model is able to recognize the protein as an IDP, with a very low percentage of secondary structure, and no distinct contacts others than the robust interaction between residues ∼50 and ∼120, in good agreement again with experimental information.[41]

Protein−Protein Complexes. Finally, we tested the ability of our simulation procedure to recognize experimental structures of protein−protein complexes. We used here as test set the strong binding complexes of the Weng's protein−protein docking benchmark 4.0.[42] Following the standard criterion we considered as strong binding cases those complexes with a binding free energy $\Delta G < -10$ kcal/mol[43] (the complexes of the test set are listed in Table S3 of the Supporting Information). We evaluated the ability of the force field to distinguish experimental complexes from false positive structures generated by protein−protein docking calculations. We want to stress that, instead of refining protein docking poses,[44] we just want to use the PACSAB simulations as a filter to discard nonnative docking poses. Thus, for each complex in the test set we performed a 1 ns DMD simulation of the experimental structure and the best scored false positive docking pose generated in a previous study.[44] We found that dissociation happened in the first few picoseconds of the trajectory, so 1 ns simulations were long enough to filter the docking poses, that had been scored using pyDock,[45] a state-of-the-art scoring funtion for protein−protein docking.[46]

In average 85% of experimental structures remained stable during the simulations (see Figure 9A), while 80% of the false positives deviated significantly from its initial conformation,

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

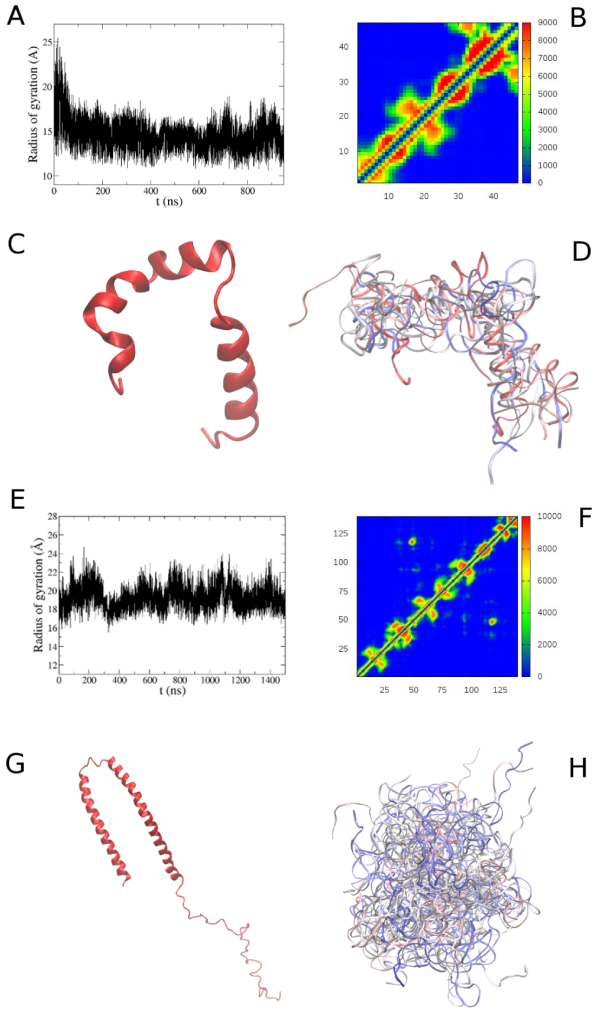


Figure 8. Structural data on ACTR (A–D) and α-synuclein (E–H). (A, E) Radius of gyration during the trajectory. (B, F) Contact map. The axes are the residue index. The color scale is from blue to red (arbitrary units). Contacts between consecutive residues have been neglected. (C) Structure of ACTR when bound to its macromolecular partner (starting structure for the DMD simulations). (G) Structure of α-synuclein when embedded in a lipid environment. (D, H) Ensemble of structures obtained during the simulation. A time-increasing colorscale (from red to blue) has been used here.

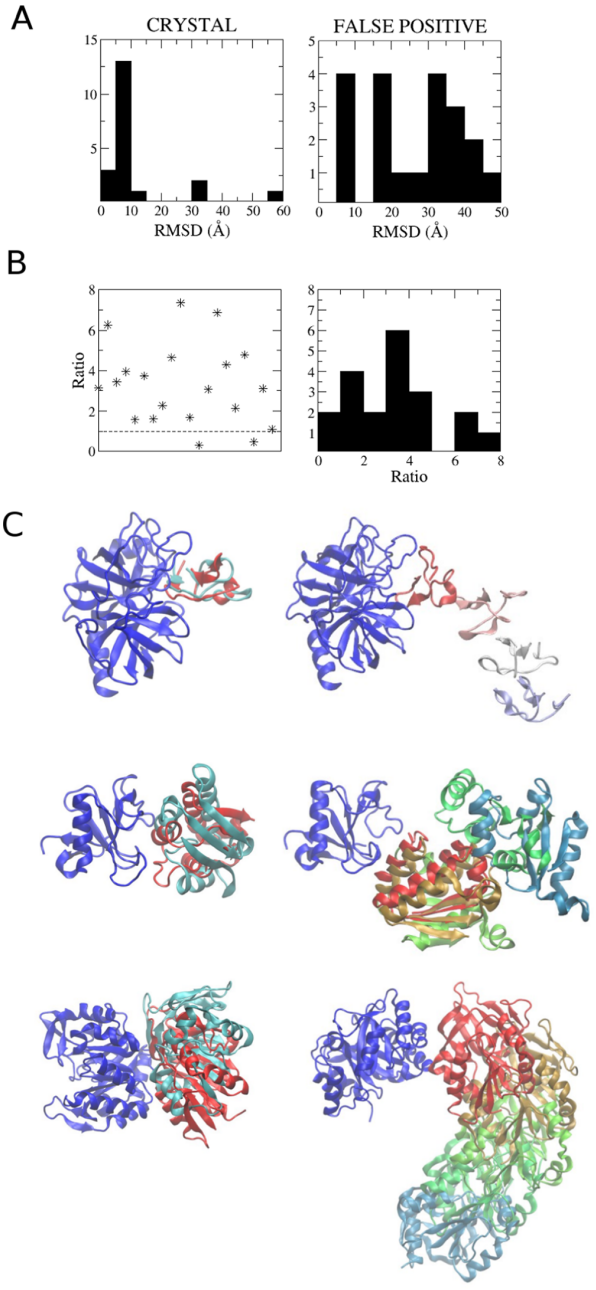


Figure 9. (A) Histogram of $RMSD_{exp}$, the RMSD with respect to the initial structure for the experimental complex (left), and histogram of $RMSD_{fp}$, the RMSD with respect to the initial structure for the best scored false positive (right). The RMSDs are calculated after a DMD simulation of 1 ns. (B) Values of the ratio $RMSD_{fp}/RMSD_{exp}$. At the left figure, symbols above the dashed line correspond to complexes for which $RMSD_{fp} > RMSD_{exp}$. At right is shown the corresponding histogram. (C) Structure of the crystal (left) and best scored false positive docking pose (right) for the complexes 1PPE, 1AY7, and 1GPW (from top to bottom). The receptor is colored in dark blue and the ligand in red. At left is shown position of the ligand (cyan) after a simulation of 1 ns; at right is shown the movement of the ligand at the beginning of the trajectory (in the frame of reference of the receptor). Several snapshots in a time-increasing colorscale (from red to blue) are shown for the ligand.

many of them leading to a complete disruption of the ligand–receptor complex (Figure 9C). In 90% of the complexes the RMS deviation from the starting structure is higher for the best scored false positive than for the experimental structure (see histogram in Figure 9B). Therefore, we have found that despite the lack of specific parametrization or the use of statistical potentials our simple DMD-based method is able not only to maintain the geometry of experimental protein–protein complexes, but to identify incorrect structures, even those that are given a strongly attractive interaction energy in docking calculations. The ability of the method to keep stable experimental complex structures while producing dissociation of nonbinding ligand–receptor orientations suggests us that the method could give good results in cross-docking[47] of proteins for which experimental information about possible binding is not available.

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929–5938

IV. CONCLUSIONS

We have constructed a physics-based discretized coarse-grained force field to represent the conformational space of proteins in solution, but also aggregated and complexed with other proteins. The force-field is implemented in a highly efficient discrete molecular dynamics algorithm which allowed us inexpensive simulations in huge systems, which would be inaccessible to standard atomistic molecular dynamics simulations. Exhaustive testing of the method shows that it is able to reproduce correctly the stability of both structured and intrinsically disordered proteins, to reproduce properly aggregation of β-amyloid peptides, and to recognize the correct structure of protein−protein complexes when compared with alternative ligand−receptor orientations which where highly scored by state-of-the-art protein−protein docking algorithms. To our knowledge, this is the first coarse-grained model able to represent both the conformational variability and interactions of proteins, including association, dissociation, and aggregation.

APPENDIX A

The index of packing n of particle i, used in the calculation of the factor γ

$$\gamma_i(n) = \frac{1}{1 + \exp((n - \alpha)/\beta)}$$

is computed as the number of faces in a truncated cube centered on particle i such that its center is near to any other particle j. The maximum value is $n = 14$, the total number of faces. $n = 14$ would correspond to a completely buried particle, $n = 0$ to a completely isolated particle. We have fitted $\alpha = 10$, the n value at which γ changes from the exposed particle ($\gamma \approx 1$) to the buried paticle ($\gamma \approx 0$) value (see Figure 10)

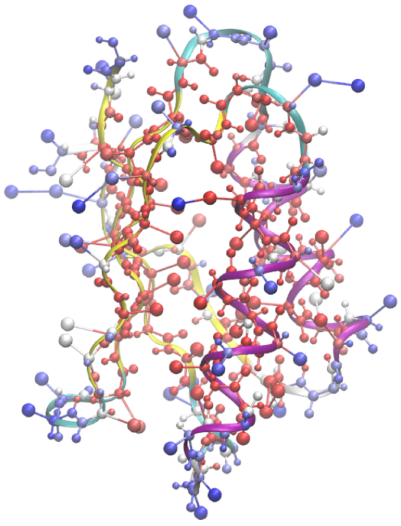


Figure 10. Structure of protein 1FVQ where particles are given a color scaled according to the value of γ (blue: exposed; red: buried).

APPENDIX B

We assume that van der Waals term of the interaction between beads A and B at the distance $r = R_{AB}$, such that V_{AB}^{vdW} has its energy minimum, is equal to the sum of atomistic van der Waals interactions at $r = R_{AB}$. The atomistic van der Waals interaction between atoms i and j is

$$V_{ij}^{at}(r) = \varepsilon_{ij}^{*}\left[\left(\frac{R_{ij}^{*}}{r}\right)^{12} - 2\left(\frac{R_{ij}^{*}}{r}\right)^{6}\right]$$

being r the distance between atom i and atom j. $\varepsilon_{ij}^{*} = \sqrt{\varepsilon_i^{*}\varepsilon_j^{*}}$ and $R_{ij}^{*} = R_i^{*} + R_j^{*}$. Supposing that all the atoms have the same van der Waals radii R_0^{*}, $R_{ij}^{*} \approx 2R_0^{*}$. Thus

$$R_A^{*} = \rho\left(\sum_i R_i^{*3}\right)^{1/3} \approx \rho(N_A R_0^{*3})^{1/3}$$

and

$$\begin{aligned} R_{AB}^{*} &= R_A^{*} + R_B^{*} \\ &\approx \rho(N_A^{1/3} + N_B^{1/3})R_0^{*} \\ &= \rho(N_A^{1/3} + N_B^{1/3})R_{ij}^{*}/2 \end{aligned}$$

where N_A (N_B) is the number of atoms included by bead A (B).

Therefore, the value of the atomistic van der Waals interacion between atoms i and j at the distance R_{AB}^{*} is

$$\begin{aligned} V_{ij}^{at}(R_{AB}^{*}) &= \varepsilon_{ij}^{*}\left[\left(\frac{R_{ij}^{*}}{R_{AB}^{*}}\right)^{12} - 2\left(\frac{R_{ij}^{*}}{R_{AB}^{*}}\right)^{6}\right] \\ &= \varepsilon_{ij}^{*}\left[\left(\frac{2/\rho}{N_A^{1/3} + N_B^{1/3}}\right)^{12} - 2\left(\frac{2/\rho}{N_A^{1/3} + N_B^{1/3}}\right)^{6}\right] \end{aligned}$$

The value of the van der Waals term of the coarse-grained potential at the distance R_{AB}^{*} is the sum of the terms corresponding to the interactions between all the atoms included in bead A and bead B:

$$V(R_{AB}^{*}) = \sum_{i\in A}\sum_{j\in B} V_{ij}^{at}(R_{AB}^{*}) = \sum_{i\in A}\sum_{j\in B} \varepsilon_{ij}^{*}(x^{12} - 2x^{6})$$

where we have defined $x = 2/[(N_A^{1/3} + N_B^{1/3})\rho]$.

Taking into account $\varepsilon_{ij}^{*} = \sqrt{\varepsilon_i^{*}\varepsilon_j^{*}}$ and assuming $\sum_{i\in A}\sum_{j\in B}\sqrt{\varepsilon_i^{*}\varepsilon_j^{*}} \approx \sqrt{(\sum_{i\in A}\varepsilon_i^{*})(\sum_{j\in B}\varepsilon_j^{*})}$ one finally obtains

$$V_{AB}^{vdW}(R_{AB}^{*}) = -\varepsilon_{AB}^{*} = \sqrt{\left(\sum_{i\in A}\varepsilon_i^{*}\right)\left(\sum_{j\in B}\varepsilon_j^{*}\right)}\,(x^{12} - 2x^{6})$$

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00660.

Details on the atomistic force fields used to constrcut the PACSAB force field; tables describing the protein−protein complexes used, simulation speed for several systems and the backbone RMSD after PACSAB simulation for proteins tested with other coarse-grained models; figures for the secondary structure changes during a trajectory for Aβ40, secondary structure propensities for monomeric and oligomeric Aβ40, intramolecular and intermolecular contact maps for Aβ40, folding trajectories of EK peptide and a β-sheet

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

peptide, Ramachandran plots of the structures after the PACSAB simulations, and contact maps and structures of the folded proteins in the benchmark after the simulations, compared to the data obtained from experimental structures (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: agusti.emperador@irbbarcelona.org (A.E.).
*E-mail: modesto.orozco@irbbarcelona.org (M.O.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Natalia Carulla for enlightening comments about the aggregation process of β-amyloid peptides and Juan Fernandez-Recio for useful comments and information on binding energies of protein complexes. We thank the Spanish Ministry of Science (Grants BIO2012-32868), the Instituto de Salud Carlos III (INB), and the European Research Council (ERC-simDNA) for support. M.O. is an ICREA Academia Fellow.

REFERENCES

(1) Morriss-Andrews, A.; Shea, J. E. Computational studies of protein aggregation: methods and applications. *Annu. Rev. Phys. Chem.* **2015**, *66*, 643−666.
(2) Orozco, M. A theoretical view of protein dynamics. *Chem. Soc. Rev.* **2014**, *43*, 5051−5066.
(3) Kleinjung, J.; Fraternali, F. Design and application of implicit solvent models in biomolecular simulations. *Curr. Opin. Struct. Biol.* **2014**, *25*, 126−134.
(4) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **2013**, *42*, 73−93.
(5) Pasi, M.; Lavery, R.; Ceres, N. PaLaCe: a coarse-grain protein model for studying mechanical properties. *J. Chem. Theory Comput.* **2013**, *9*, 785−793.
(6) Hills, R. D., Jr.; Lu, L.; Voth, G. A. Multiscale coarse-graining of the protein energy landscape. *PLoS Comput. Biol.* **2010**, *6*, e1000827.
(7) Kar, P.; Gopal, S. M.; Cheng, Y. M.; Predeus, A.; Feig, M. RIMO: A Transferable Coarse-grained Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 3769−3788.
(8) Kapoor, A.; Travesset, A. Folding and stability of helical bundle proteins from coarse-grained models. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 1200−1211.
(9) Bereau, T.; Deserno, M. Generic doarse-grained model for protein folding and aggregation. *J. Chem. Phys.* **2009**, *130*, 235106.
(10) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Oldziej, S.; Wachucik, K.; Scheraga, H. A. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B* **2007**, *111*, 260−285.
(11) Sterpone, F.; Melchionna, S.; Tuffery, P.; Pasquali, S.; Mousseau, N.; Cragnolini, T.; Chebaro, Y.; St-Pierre, J.-F.; Kalimeri, M.; Barducci, A.; Laurin, Y.; Tek, A.; Baaden, M.; Nguyen, P.-H.; Derreumaux, P. The OPEP protein model: from single molecule-s,amyloid formation, crowding and hydrodynamics to DNA/RNA systems. *Chem. Soc. Rev.* **2014**, *43*, 4871−4893.
(12) Chebaro, Y.; Pasquali, S.; Derreumaux, P. The coase-grained OPEP force field for non-amyloid and amyloid proteins. *J. Phys. Chem. B* **2012**, *116*, 8741−8752.
(13) Auer, S. Phase diagram of polypeptide chains. *J. Chem. Phys.* **2011**, *135*, 175103.
(14) Urbanc, B.; Cruz, L.; Yun, S.; Buldyrev, S. V.; Bitan, G.; Teplow, D. B.; Stanley, H. E. In silico study of amyloid beta-protein folding and oligomerization. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17345−17350.
(15) Orozco, M.; Orellana, L.; Hospital, A.; Naganathan, A. N.; Emperador, A.; Carrillo, O.; Gelpi, J. L. Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings. *Adv. Protein Chem. Struct. Biol.* **2011**, *85*, 183−215.
(16) Emperador, A.; Meyer, T.; Orozco, M. Protein flexibility from discrete molecular dynamics simulations using quasi-physical potentials. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 83−94.
(17) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B* **2004**, *108*, 750−760.
(18) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111*, 7812−7824.
(19) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.
(20) Anandakrishnan, R.; Drozdetski, A.; Walker, R. C.; Onufriev, A. V. Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. *Biophys. J.* **2015**, *108*, 1153−1164.
(21) Emperador, A.; Meyer, T.; Orozco, M. United-Atom Discrete Molecular Dynamics of Proteins Using Physics-Based Potentials. *J. Chem. Theory Comput.* **2008**, *4*, 2001−2010.
(22) Ding, F.; Tsao, D.; Nie, H.; Dokholyan, N. V. Ab initio folding of proteins with all-atom discrete molecular dynamics. *Structure* **2008**, *16*, 1010−1018.
(23) Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **1996**, *105*, 1902−1921.
(24) Lazaridis, T.; Karplus, M. Effective energy function for proteins in solution. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133−152.
(25) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *J. Chem. Phys.* **2014**, *140*, 224104.
(26) Miyazawa, S.; Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **1996**, *256*, 623−644.
(27) Zhang, C.; Vasmatzis, G.; Cornette, J. L.; DeLisi, C. Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.* **1997**, *267*, 707−26.
(28) Bitan, G.; Vollers, S. S.; Teplow, D. B. Elucidation of primary structure elements controlling early amyloid beta-protein oligomerization. *J. Biol. Chem.* **2003**, *278*, 34882−34889.
(29) Olubiyi, O. O.; Strodel, B. Structures of the amyloid β-peptides Aβ1−40 and Aβ1−42 as influenced by pH and a D-peptide. *J. Phys. Chem. B* **2012**, *116*, 3280−3291.
(30) Nasica-Labouze, J.; Nguyen, H.; Sterpone, F.; Berthomieu, O.; Buchete, N.-V.; Cote, S.; De Simone, A.; Doig, A. J.; Faller, P.; Garcia, A.; Laio, A.; Li, M. S.; Melchionna, S.; Mousseau, N.; Mu, Y.; Paravastu, A.; Pasquali, S.; Rosenman, D. J.; Strodel, B.; Tarus, B.; Viles, J. H.; Zhang, T.; Wang, Ch.; Derreumaux, P. Amyloid β protein and Alzheimer's disease: when computer simulations complement experimental studies. *Chem. Rev.* **2015**, *115*, 3518−3563.
(31) Takeda, T.; Klimov, K. Probing the effect of amino-terminal truncation for Aβ40 peptides. *J. Phys. Chem. B* **2009**, *113*, 6692−6702.
(32) Kim, S.; Takeda, T.; Klimov, K. Mapping conformational ensembles of Aβ oligomers in molecyular dynamics simulations. *Biophys. J.* **2010**, *99*, 1949−1958.
(33) Lin, Y.-S.; Bowman, G. R.; Beauchamp, K. A.; Pande, V. S. Investigating how peptide length and a pathogenic mutation modify the structural ensemble of amyloid beta monomer. *Biophys. J.* **2012**, *102*, 315−324.
(34) Pujol-Pina, R.; Vilaprinyo-Pascual, S.; Mazzucato, R.; Arcellla, A.; Vilaseca, M.; Orozco, M.; Carulla, N. *Sci. Rep.* **2015**, *5*, 14809.
(35) Barz, B.; Olubiyi, O. O.; Strodel, B. Early amyloid β-protein aggregation precedes conformational change. *Chem. Commun.* **2014**, *50*, 5373−5375.

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

(36) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Perez, A.; Camps, J.; Hospital, A.; Gelpi, J. L.; Orozco, M. A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 796−801.
(37) Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Perez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpi, J. L.; Orozco, M. MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure* **2010**, *18*, 1399−1409.
(38) Candotti, M.; Perez, A.; Ferrer-Costa, C.; Rueda, M.; Meyer, T.; Gelpi, J. L.; Orozco, M. Exploring early stages of the chemical unfolding of proteins at the proteome scale. *PLoS Comput. Biol.* **2013**, *9*, e1003393.
(39) Demarest, S. J.; Martinez-Yamout, M.; Chung, J.; Chen, H.; Xu, W.; Dyson, H. J.; Evans, R. M.; Wright, P. E. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **2002**, *415*, 549−553.
(40) Kjaergaard, M.; Norholm, A. B.; Hendus-Altenburger, R.; Pedersen, S. F.; Poulsen, F. M.; Kragelund, B. B. Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II? *Protein Sci.* **2010**, *19*, 1555−1564.
(41) Esteban-Martin, S.; Silvestre-Ryan, J.; Bertoncini, C. W.; Salvatella, X. Identification of fibril-like tertiary contacts in soluble monomeric α-synuclein. *Biophys. J.* **2013**, *105*, 1192−1198.
(42) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein-protein docking benchmark version 4.0. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 3111−3114.
(43) Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. A.; Bonvin, A. M.; Janin, J. A structure-based benchmark for protein-protein binding affinity. *Protein Sci.* **2011**, *20*, 482−491.
(44) Emperador, A.; Solernou, A.; Sfriso, P.; Pons, C.; Gelpi, J. L.; Fernandez-Recio, J.; Orozco, M. Efficient relaxation of protein-protein interfaces by discrete molecular dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 1222−1229.
(45) Cheng, T. M.; Blundell, T. L.; Fernandez-Recio, J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins: Struct., Funct., Genet.* **2007**, *68*, 503−515.
(46) Fernandez-Recio, J.; Totrov, M.; Abagyan, R. Soft protein-protein docking in internal coordinates. *Protein Sci.* **2002**, *11*, 280−291.
(47) Sacquin-Mora, S.; Carbone, A.; Lavery, R. Identification of protein interaction partners and protein-protein interaction sites. *J. Mol. Biol.* **2008**, *382*, 1276−1289.

DOI: 10.1021/acs.jctc.5b00660
J. Chem. Theory Comput. 2015, 11, 5929−5938

7.4 Discrete Molecular Dynamics: A Review

Title: **Discrete Molecular Dynamics: Foundations and Biomolecular Applications**

Authors: Pedro Sfriso*, Agustí Emperador*, Josep Lluís Gelpí, and Modesto Orozco

Stage: Published

Book: Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods

Type: Review. Book Chapter.

Author Contribution: P. S. contributed with ideas and to the writing of the chapter.

Citation: Computational Approaches to Protein DynamicsFrom Quantum to Coarse-Grained Methods, Edited by Monika Fuxreiter, CRC Press 2015, Pages 339–362

Summary

This is an extensive review where we discuss the state of the art for biological applications of Discrete Molecular Dynamics.

Content: Dynamic Nature of Proteins, Basic dMD Algorithm, dMD Force Fields, Implementation of dMD, dMD in Web Servers, Application of dMD in Biomolecular Problems, Protein Folding, Protein Structural Refinement and Design, RNA Structural Predictions, Protein Aggregation, Equilibrium Protein Dynamics, Conformational Transitions and Protein–Protein Docking.

Chapter 8: Discussion and Concluding Remarks

This chapter is structured, according to mandatory guidelines, in three sections. In the first one, a schematic summary of findings in this Thesis is presented. In the second part, I discuss the results and finally the general conclusions are stated.

8. 1 Summary of Findings in this Thesis

General

We presented methodological developments leading to an improvement in our capacities to sample conformational transitions in proteins. Our methods populate paths joining protein conformations in an efficient and extensible manner, but with low-resolution structures. In a second step, we prepared those structures for more detailed simulations, where atomistic detail can be recovered. We show how, by incorporating very low resolution experimental data, such as coevolutionary signals, the analysis of conformational pathways in protein can be simplified, allowing us to outline transition routes that agree with known experimental information.

Methodological advances

We obtained algorithms to speed up the computation of conformational transitions in proteins.
We developed a protocol to predict conformational transitions when one conformation is known provided that enough homologous sequences are available.
We extend Partial Least Square projections to automatically derive collective variables for any conformational transition.
We propose a strategy to parameterize SBM to maximize the coincidence with any external signal.

Technical problems addressed

We extended the Maxwell-Demon sampling strategy to bias trajectories smoothly.
We implemented multiple minima energy potentials in discrete Molecular Dynamics.
We created structure-based models for capturing from 1 to 500 (redundant) reference structures
We developed algorithms to enhance sampling consistently with 1 to 500 *known* protein structures.
We designed novel statistics to couple noisy data to simulations. We successfully filtered noisy coevolution contacts.

We designed a database with +60000 simulations of conformational transitions. We automatically classified them.

Make the tools available for the community

We provided the community with on-line tools through our web servers: http://mmb.irbbarcelona.org/MDdMD/ and http://mmb.irbbarcelona.org/GOdMD/. Our database of protein motions can be found at http:// mmb.irbbarcelona.org/TransAtlas/.

8.2 General Discussion

In this final part of this Thesis, I will briefly discuss the main results obtained and their relevance in the field of protein dynamics: from the methodological foundations (Chapter 3 and 4), to the eventual predictions of conformational transitions in proteins (Chapter 5) or the extension of coarse-grained sampling strategy to large datasets (Chapter 6).

The study of protein flexibility needs and probably will need in the near future computer simulations. Even though atomistic MD has proven its value, sampling issues are limiting its ability to provide a proteome-scale picture of the conformational space of proteins. The design of new, lower resolution, but computationally faster methods is required. Along this Thesis we aimed to develop computational tools as part of a novel strategy to simulate biological processes, mainly conformational transitions. Inspired by previous works, we propose a multi-step approach where initial CG simulations lead a coarse exploration of the conformational space and leave the dissection of fine detail only where it matters, saving hours of computational time.

In this Thesis, we presented two methods to trace conformational transitions when two end points are known. The first one, **MDdMD**, is based on physical potentials, explicitly models all protein atoms but hydrogen. It completes transition paths in hours on a single processor, being ideal for exploratory simulations. The main application of MDdMD is to find transitions paths and populate them with reasonable conformers that are in turn, the starting point for higher-level calculations. The sampling will be automatically enhanced since several starting points prevent simulations from being trapped in the initial energy basin. Several methods naturally benefit from MDdMD strategy, being complementary to Milestoning (*215*, *219*), Metadynamics (*223*, *226*), targeted MD (*199*), forward flux sampling (*252*) or other renamed methods. Our second method, **GOdMD**, replaces physical potentials with Go-like potentials and, at the same time, reduces the amino acid representation to a single bead. Notably, we gained efficiency and accuracy in describing transition paths: known transition intermediates are visited to a greater extent than in MDdMD, showing that protein topology dictates most of the conformational motion (*114*, *166*, *253*). The success of structural based potentials indicates that transitions paths were tuned to avoid massive change in the protein contacts, opening the possibility to model conformational transitions with very low resolution for proteins. The major drawback is that we lose the immediate coupling to atomistic MD with the Cα resolution.

Coarse-graining speeds up calculations reducing the effective barriers separating distinct conformers, boosting the diffusive motion through configuration space In other words: CG accelerates diffusive properties of 'some' events. But attention is needed when interpreting CG dynamics since we cannot guarantee that all processes are accelerated to the same extend (*254*). Similarly, any correspondence between CG and real transition times should be fortuitous.

We implemented a Go-like model in GOdMD method to describe the energy landscape with smooth funnel-shaped surface, aiming for efficiency. We needed the efficiency for high-throughput studies, like for instance, to test the impact on dynamics of disease-related mutations (or any other kind of contact-like information) in a combinatorial fashion. By combination of low number of particles, simple energy functions and dMD, transitions path are very efficiently obtained. In fact, our method can trace in one laptop processor a conformational transition with an average time of two minutes, meaning that with supercomputer resources proteome-scale simulations are possible.

Contacts not present in the native structure, but present in alternative conformations, can be captured by coevolution analysis of the protein sequence. Coevolution signal is typically weak and is mostly concentrated on top ranked coevolution contacts, which unfortunately are little informative for dynamics. Consequently, to investigate dynamic motions impressed in sequence, thousands contacts must be considered to evaluate their information-load. Here is when fast CG models played a decisive role. We discerned relevant contacts by assaying them with individual trajectories, and once identified, we gather them to predict consistent transition paths. Notably, despite the combinatorial nature of our approach, the automated protocol converges in the hour time-range (again in a laptop computer). Enriching dynamics with coevolution allowed us to overcome (when such data is available), the limitation of knowing both ends of the conformational transition. Plus, on the other side, the analysis of coevolution data with CG models concluded that coevolution signal extends almost 10-fold further, in ranked contacts, than initially estimated (*255*).

Trying to move our methods to the proteome scale to give a broader picture of protein conformational pathways, we pre-computed nearly all-possible conformational transitions between structures deposited in the Protein Data Bank. Apart from the transition trajectory, we aim to learn from the conformational changes observed. We extracted biophysical data regarding the change in shape that we used to classify the motions, hoping that predicted paths are useful to test ideas or design experimental set-ups. One example where we used this classification is to show that the depth of the coevolution information needed to successfully predict a protein conformer depends of the type of motion (Chapter 5).

After an initial exploration of the conformational space, the next step is to extend the analysis with other methods addressing fine conformational details. In an effort for a multiscale approach, we reconstructed back the atomistic models from the Cα trajectory. In doing so, we provide with 10-50 intermediate snapshots per conformational transition ready to be the starting point of a standard (or biased) MD simulations. For those users with less expertise, or for those interested in automated screenings, we linked our method to our MDWEB (*251*) platform that prepares all input files for major simulation packages automatically.

As discussed above MD often needs to be complemented with enhanced sampled techniques to study large-scale dynamics. A common requirement to such techniques is the detection of collective variables that capture the conformational change, something not trivial in the absence of a previous knowledge on the conformational pathway (*256*). Targeting the molecular simulation community, we derived an automated method to identify the collective variables of each CG conformational transition (Chapter 6). To this end, we adapted the statistical sound Partial Least Square regression technique to determine which set of internal distances capture most of the variance displayed at the conformational change. These variables are reasonable choices, for example, in Umbrella Sampling or Metadynamic simulations. The result of our exploration leads to +60000 independent transition trajectories, publically available at our TransAtlas database.

Finally, it has to be said that no alternative method has defeat yet MD as the dominant method to study protein dynamics. However, CG methods will steadily be more protagonists, as better specific-purpose algorithms will appear. Also, new parameterization procedures based directly on experiments will undoubtedly play a central role, accelerating the simulation-experiment feedback loop, particularly during the model construction process. In one line, we should not forget about simpler models to uncover the fascinating nature of macromolecular motion.

Future Challenges

Computational Biology became a mature field. Equilibrium dynamics of proteins and prediction of small peptides structure is nowadays customary. There are several processes where progress in the next decade is expected, for instance ligand docking or folding of small proteins. But there are others that will require inventive strategies, for example predicting dynamics of intrinsically disordered proteins (IDP). IDPs break the paradigm of structure to function because they lack from an ordered three-dimensional structure (although they can occasionally structure at binding to other macromolecules). No effective tool for simulating IDPs dynamics exits since, in one hand, atomistic MD simulation are unaffordable, and in other hand, more reliable CG models need at some point a reference structure. Another example where innovative approaches are needed is the protein-protein recognition studies, especially those mediated by no-purely amino acidic contacts.

8.3 General Conclusions

In this Thesis I worked to expand the applicability of simplified methods focusing on conformational transitions. I aimed to contribute to bridge gap between computation and biology and start to face the ubiquitous sampling problem. I collected the following conclusions:

- The path connecting different conformers of proteins can be outlined from simple coarse-grained simulation methods. With careful design and iterative approaches the resolution and accuracy can be increased (Chapter 3 and 4).
- We obtained a hybrid protocol to predict functional conformational space of proteins. (Chapter 5).
- We observed that step potentials and discrete Molecular Dynamics excel on dealing with uncertain experimental (or bioinformatic) data. Flat potentials are a natural way of handling noisy signals (Chapter 5).
- Coevolution information goes far beyond the folding contacts, allowing the characterization of protein landscapes when incorporated into efficient sampling algorithms (Chapter 5).
- The speed of the developed technology allowed us to study in a proteome-scale the conformational transition landscape of proteins (Chapter 6).

References

1. A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino, A. R. Ortíz, An Analysis of Core Deformations in Protein Superfamilies. *Biophys. J.* **88**, 1291–1299 (2005).

2. M. Orozco, The dynamic view of proteins: Comment on "Comparing proteins to their internal dynamics: Exploring structure–function relationships beyond static structural alignments." *Phys Life Rev* **10**, 29–30 (2012).

3. C. Micheletti, Comparing proteins by their internal dynamics: exploring structure-function relationships beyond static structural alignments. *Phys Life Rev* **10**, 1–26 (2013).

4. H. G. Dos Santos, J. Klett, R. Mendez, U. Bastolla, Biochimica et Biophysica Acta. *Biochim. Biophys. Acta* **1834**, 836–846 (2013).

5. K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).

6. E. Z. Eisenmesser, D. A. Bosco, M. Akke, D. Kern, Enzyme dynamics during catalysis. *Science* **295**, 1520–1523 (2002).

7. H. Frauenfelder, S. G. Sligar, P. G. Wolynes, The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).

8. A. Bejan, S. Lorente, The constructal law of design and evolution in nature. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **365**, 1335–1347 (2010).

9. J. A. Velazquez-Muriel *et al.*, Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Struct. Biol. [Online]* **9**, 6 (2009).

10. H. M. Berman *et al.*, The protein data bank. *Nuc. Acids Res.* **28**, 235–242 (2000).

11. K. Wüthrich, Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science* **243**, 45–50 (1989).

12. A. Mittermaier, L. E. Kay, New tools provide new insights in NMR studies of protein dynamics. *Science* **312**, 224–228 (2006).

13. P. Schanda, B. Brutscher, Very Fast Two-Dimensional NMR Spectroscopy for Real-Time Investigation of Dynamic Events in Proteins on the Time Scale of Seconds. *J. Am. Chem. Soc.* **127**, 8014–8015 (2005).

14. C. Charlier *et al.*, Nanosecond Time Scale Motions in Proteins Revealed by High-Resolution NMR Relaxometry. *J. Am. Chem. Soc.* **135**, 18665–18672 (2013).

15. F. Schotte *et al.*, Watching a signaling protein function in real time via 100-ps time-resolved Laue crystallography. *Proc. Natl. Acad. Sci. USA* **109**, 19256–19261 (2012).

16. H. N. Chapman *et al.*, Femtosecond X-ray protein nanocrystallography. *Nature* **470**, 73–77 (2011).

17. R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, J. Hajdu, Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**, 752–757 (2000).

18. T. Heyduk, Measuring protein conformational changes by FRET/LRET. *Curr Opin Biotechnol* **13**, 292–296 (2002).

19. A. T. Brunger, P. Strop, M. Vrljic, S. Chu, K. R. Weninger, Three-dimensional molecular modeling with single molecule FRET. *Journal of Structural Biology* **173**, 497–505 (2011).

20. P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, D. I. Svergun, Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **129**, 5656–5664 (2007).

21. C. D. Putnam, M. Hammel, G. L. Hura, J. A. Tainer, X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40**, 191–285 (2007).

22. P. Cossio, G. Hummer, Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *Journal of Structural Biology* **184**, 427–437 (2013).

23. N. Fischer, A. L. Konevega, W. Wintermeyer, M. V. Rodnina, H. Stark, Ribosome dynamics and tRNA movement by

time-resolved electron cryomicroscopy. *Nature* **466**, 329–333 (2010).

24. G. Di Fede *et al.*, Structure of the anaphase-promoting complex/cyclosome interacting with a mitotic checkpoint complex. *Science* **323**, 1477–1481 (2009).

25. M. Levitt, The birth of computational structural biology. *Nat. Struct Biol.* **8**, 392–393 (2001).

26. M. Karplus, *Molecular dynamics of biological macromolecules: a brief history and perspective.* (Biopolymers, 2003), pp. 350–358.

27. B. J. Alder, T. E. Wainwright, Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **27**, 1208–1209 (1957).

28. A. Rahman, F. H. Stillinger, Molecular Dynamics Study of Liquid Water. *J. Chem. Phys.* **55**, 3336–3359 (1971).

29. J. A. McCammon, B. R. Gelin, M. Karplus, Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).

30. M. P. Allen, D. J. Tildesley, Computer Simulation of Liquids. *Oxford University Press* (1989).

31. D. C. Rapaport, The Art of Molecular Dynamics Simulation. (2004).

32. M. Bixon, S. Lifson, Potential functions and conformations in cycloalkanes. *Tetrahedron* **23**, 769–784 (1967).

33. S. Lifson, M. Levitt, On obtaining energy parameters from crystal structure data. *Computers & Chemistry* **3**, 49–50 (1979).

34. M. Levitt, S. Lifson, Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* **46**, 269–279 (1969).

35. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).

36. D. M. Zuckerman, Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.* **40**, 41–62 (2011).

37. R. Car, M. Parrinello, Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55**, 2471–2474 (1985).

38. M. Orozco, A theoretical view of protein dynamics. *Chemical Society Reviews* **43**, 5051–5066 (2014).

39. H. M. Senn, W. Thiel, QM/MM studies of enzymes. *Current opinion in chemical biology* **11**, 182–187 (2007).

40. R. Das, D. Baker, Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008).

41. Y. Ding, A. B. Mamonov, D. M. Zuckerman, Efficient equilibrium sampling of all-atom peptides using library-based Monte Carlo. *J. Phys. Chem. B* **114**, 5870–5877 (2010).

42. Z. Li, H. A. Scheraga, Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. USA* **84**, 6611–6615 (1987).

43. N. Kantarci-Carsibasi, T. Haliloglu, P. Doruker, Conformational Transition Pathways Explored by Monte Carlo Simulation Integrated with Collective Modes☆. *Biophys. J.* **95**, 5862–5873 (2008).

44. S. L. Seyler, O. Beckstein, Sampling large conformational transitions: adenylate kinase as a testing ground. *Molecular Simulation*, 1–23 (2014).

45. O. Beckstein, E. J. Denning, J. R. Perilla, T. B. Woolf, Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open↔ closed transitions. *J. Mol. Biol.* **394**, 160–176 (2009).

46. J. R. Perilla, O. Beckstein, E. J. Denning, T. B. Woolf, Computing ensembles of transitions from stable states: Dynamic importance sampling. *Journal of Computational Chemistry* **32**, 196–209 (2010).

47. M. Orozco *et al.*, Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings. *Adv Protein Chem Struct Biol* **85**, 183–215 (2011).

48. A. Emperador, O. Carrillo, M. Rueda, M. Orozco, Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophys. J.* **95**, 2127–2138 (2008).

49. M. Levitt, C. Sander, P. S. Stern, Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* **181**, 423–447 (1985).

50. N. Go, T. Noguti, T. Nishikawa, Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* **80**, 3696–3700 (1983).

51. B. Brooks, M. Karplus, Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA* **80**, 6571–6575 (1983).

52. R. Elber, M. Karplus, Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science* **235**, 318–321 (1987).

53. L. Orellana *et al.*, Approaching elastic network models to molecular dynamics flexibility. *Journal of Chemical Theory and Computation* **6**, 2910–2923 (2010).

54. J. R. Lopez-Blanco, J. I. Garzón, P. Chacón, iMod: multipurpose normal mode analysis in internal coordinates. *Bioinformatics* **27**, 2843–2850 (2011).

55. M. Bathe, A finite element framework for computation of protein normal modes and mechanical response. *Proteins: Struct., Funct., Bionf.* **70**, 1595–1609 (2007).

56. E. A. Proctor, F. Ding, N. V. Dokholyan, Discrete molecular dynamics. *WIREs Comput. Mol. Sci.* **1**, 80–92 (2011).

57. P. Sfriso, A. Emperador, J. Gelpí, M. Orozco, in *Series in Computational Biophysics*, (CRC Press, 2014), pp. 339–362.

58. Y. Zhou, M. Karplus, Folding thermodynamics of a model three-helix-bundle protein. *Proc. Natl. Acad. Sci. USA* **94**, 14429–14432 (1997).

59. Y. Zhou, M. Karplus, Interpreting the folding kinetics of helical proteins. *Nature* **401**, 400–403 (1999).

60. F. Ding *et al.*, Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA* **14**, 1164–1173 (2008).

61. F. Ding, C. A. Lavender, K. M. Weeks, N. V. Dokholyan, Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat. Methods* **9**, 603–608 (2012).

62. F. Ding, S. Buldyrev, N. Dokholyan, Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophys. J.* **88**, 147–155 (2005).

63. F. Ding, D. Tsao, H. Nie, N. V. Dokholyan, Ab initio folding of proteins with all-atom discrete molecular dynamics. *Structure* **16**, 1010–1018 (2008).

64. D. Shirvanyants, F. Ding, D. Tsao, S. Ramachandran, N. V. Dokholyan, Discrete molecular dynamics: an efficient and versatile simulation method for fine protein characterization. *J. Phys. Chem. B* **116**, 8375–8382 (2012).

65. A. Emperador, T. Meyer, M. Orozco, Protein flexibility from discrete molecular dynamics simulations using quasi-physical potentials. *Proteins: Struct., Funct., Bionf.* **78**, 83–94 (2010).

66. A. Emperador *et al.*, Efficient Relaxation of Protein–Protein Interfaces by Discrete Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* **9**, 1222–1229 (2013).

67. S. Ramachandran, P. Kota, F. Ding, N. V. Dokholyan, Automated minimization of steric clashes in protein structures. *Proteins: Struct., Funct., Bionf.* **79**, 261–270 (2011).

68. E. A. Proctor, S. Yin, A. Tropsha, N. V. Dokholyan, Discrete molecular dynamics distinguishes nativelike binding poses from decoys in difficult targets. *Biophys. J.* **102**, 144–151 (2012).

69. F. Ding, N. V. Dokholyan, Dynamical roles of metal ions and the disulfide bond in Cu, Zn superoxide dismutase folding and aggregation. *Proc. Natl. Acad. Sci. USA* **105**, 19696–19701 (2008).

70. B. Urbanc, J. Borreguero, L. Cruz, H. Stanley, Ab initio discrete molecular dynamics approach to protein folding and aggregation. *Methods Enzymol.* **412**, 314–338 (2006).

71. B. Urbanc, M. Betnel, L. Cruz, G. Bitan, D. Teplow, Elucidation of amyloid β-protein oligomerization mechanisms: discrete molecular dynamics study. *J. Am. Chem. Soc.* **132**, 4266–4280 (2010).

72. B. Urbanc *et al.*, In silico study of amyloid beta-protein folding and oligomerization. *Proc. Natl. Acad. Sci. USA* **101**, 17345–17350 (2004).

73. M. A. Khan, M. C. Herbordt, Parallel discrete molecular dynamics simulation with speculation and in-order commitment. *J. Comput. Phys.* **230**, 6563–6582 (2011).

74. B. Sukhwani, M. Chiu, M. A. Khan, M. C. Herbordt, Effective floating point applications on FPGAs: Examples from molecular modeling. *HPEC'09: Proc. of the Workshop on High Performance Embedded Computing* (2009).

75. A. Warshel, M. Levitt, S. Lifson, Consistent force field for calculation of vibrational spectra and conformations of some amides and lactam rings. *Journal of Molecular Spectroscopy* **33**, 84–99 (1970).

76. S. A. Adcock, J. A. McCammon, Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **106**, 1589–1615 (2006).

77. A. J. Mulholland, Modelling enzyme reaction mechanisms, specificity and catalysis. *Drug Discov. Today* **10**, 1393–1402 (2005).

78. H. M. Senn, W. Thiel, QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **48**, 1198–1229 (2009).

79. A. Warshel, R. M. Weiss, An empirical valence bond approach for comparing reactions in solutions and in enzymes. *J. Am. Chem. Soc.* **102**, 6218–6226 (1980).

80. S. C. L. Kamerlin, A. Warshel, The empirical valence bond model: theory and applications. *WIREs Comput. Mol. Sci.* **1**, 30–45 (2011).

81. H. I. Ingólfsson *et al.*, The power of coarse graining in biomolecular simulations. *WIREs Comput. Mol. Sci.* **4**, 225–248 (2013).

82. L. Monticelli *et al.*, The MARTINI Coarse-Grained Force Field: Extension to Proteins. *Journal of Chemical Theory and Computation* **4**, 819–834 (2008).

83. X. Periole, M. Cavalli, S.-J. Marrink, M. A. Ceruso, Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *Journal of Chemical Theory and Computation* **5**, 2531–2543 (2009).

84. A. Liwo *et al.*, A united‐residue force field for off‐lattice protein‐structure simulations. I. Functional forms and parameters of long‐range side‐chain interaction potentials from protein crystal data. *Journal of Computational Chemistry* **18**, 849–873 (1997).

85. A. Liwo, Y. He, H. A. Scheraga, Coarse-grained force field: general folding theory. *Phys Chem Chem Phys* **13**, 16890 (2011).

86. S. Ołdziej *et al.*, Protein-structure prediction using ahierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proc. Natl. Acad. Sci. USA* **102**, 7547–7552 (2005).

87. A. Liwo, M. Khalili, H. A. Scheraga, Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. USA* **102**, 2362–2367 (2005).

88. M. Pasi, R. Lavery, N. Ceres, PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties. *Journal of Chemical Theory and Computation* **9**, 785–793 (2013).

89. D. Reith, M. Pütz, F. Müller-Plathe, Deriving effective mesoscale potentials from atomistic simulations. *Journal of Computational Chemistry* **24**, 1624–1636 (2003).

90. P. Kar, S. M. Gopal, Y.-M. Cheng, A. Predeus, M. Feig, PRIMO: A Transferable Coarse-Grained Force Field for Proteins. *Journal of Chemical Theory and Computation* **9**, 3769–3788 (2013).

91. F. Ding, N. V. Dokholyan, Emergence of protein fold families through rational design. *PLoS Comput Biol* **2**, e85 (2006).

92. T. Lazaridis, M. Karplus, Effective energy function for proteins in solution. *Proteins: Struct., Funct., Bionf.* **35**, 133–152 (1999).

93. S. Yin, L. Biedermannova, J. Vondrasek, N. V. Dokholyan, MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening. *J. Chem. Inf. Model* **48**, 1656–1662 (2008).

94. L. Darré *et al.*, SIRAH: A Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics. *Journal of Chemical Theory and Computation* **11**, 723–739 (2015).

95. P. Derreumaux, N. Mousseau, Coarse-grained protein molecular dynamics simulations. *J. Chem. Phys.* **126**, 025101 (2007).

96. J. Maupetit, P. Tuffery, P. Derreumaux, A coarse-grained protein force field for folding and structure prediction. *Proteins: Struct., Funct., Bionf.* **69**, 394–408 (2007).

97. N. Basdevant, D. Borgis, T. Ha-Duong, A Coarse-Grained Protein−Protein Potential Derived from an All-Atom Force Field. *J. Phys. Chem. B* **111**, 9390–9399 (2007).

98. T. Bereau, M. Deserno, Generic coarse-grained model for protein folding and aggregation. *J. Chem. Phys.* **130**, 235106

(2009).

99. S. Izvekov, G. A. Voth, A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **109**, 2469–2473 (2005).

100. W. G. Noid *et al.*, The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **128**, 244114 (2008).

101. S. Izvekov, G. A. Voth, Multiscale coarse graining of liquid-state systems. *J. Chem. Phys.* **123**, 134105 (2005).

102. Y. Ueda, H. Taketomi, N. G, Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme. *Biopolymers* **17**, 1531–1548 (1978).

103. H. Taketomi, Y. Ueda, N. Gō, Studies on protein folding, unfolding and fluctuations by computer simulation. *Int. J. Pept. Protein Res.* **7**, 445–459 (1975).

104. J. Bryngelson, J. Onuchic, N. Socci, P. Wolynes, Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Struct., Funct., Bionf.* **21**, 167–195 (1995).

105. J. N. Onuchic, Z. Luthey-Schulten, P. G. Wolynes, Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).

106. D. E. Kim, H. Gu, D. Baker, The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci. USA* **95**, 4982–4986 (1998).

107. V. Muñoz, W. A. Eaton, A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA* **96**, 11311–11316 (1999).

108. G. Hummer, A. Szabo, Kinetics from Nonequilibrium Single-Molecule Pulling Experiments. *Biophys. J.* **85**, 5–15 (2003).

109. O. Dudko, G. Hummer, A. Szabo, Intrinsic Rates and Activation Free Energies from Single-Molecule Pulling Experiments. *Phys. Rev. Lett.* **96**, 108101 (2006).

110. J. N. Onuchic, H. Nymeyer, A. E. García, J. Chahine, N. D. Socci, in *Multiple values selected*, (Elsevier, 2000), vol. 53, pp. 87–152.

111. Y. Levy, S. S. Cho, J. N. Onuchic, P. G. Wolynes, A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. *J. Mol. Biol.* **346**, 1121–1145 (2005).

112. P. C. Whitford, O. Miyashita, Y. Levy, J. N. Onuchic, Molecular Dynamics Studies on the Conformational Transitions of Adenylate Kinase: A Computational Evidence for the Conformational Selection Mechanism. *J. Mol. Biol.* **2013**, 1661–1671 (2007).

113. R. B. Best, G. Hummer, W. A. Eaton, Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17874–17879 (2013).

114. D. Baker, A surprising simplicity to protein folding. *Nature* **405**, 39–42 (2000).

115. U. Bastolla, M. Vendruscolo, E. W. Knapp, A statistical mechanical method to optimize energy functions for protein folding. *Proc. Natl. Acad. Sci. USA* **97**, 3977–3981 (2000).

116. S. Miyazawa, R. L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

117. S. Miyazawa, R. L. Jernigan, Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644 (1996).

118. M. J. Sippl, Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct., Funct., Bionf.* **17**, 355–362 (1993).

119. D. Chivian *et al.*, Automated prediction of CASP-5 structures using the Robetta server. *Proteins: Struct., Funct., Bionf.* **53**, 524–533 (2003).

120. R. Bonneau *et al.*, De Novo Prediction of Three-dimensional Structures for Major Protein Families. *J. Mol. Biol.* **322**, 65–78 (2002).

121. C. A. Rohl, C. E. Strauss, K. M. Misura, D. Baker, Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).

122. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct., Funct., Bionf.* **37**, 171–176 (1999).

123. M. Tirion, Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **77**, 1905–1908 (1996).

124. J. A. Kovacs, P. Chacón, R. Abagyan, Predictions of protein flexibility: First-order measures. *Proteins: Struct., Funct., Bionf.* **56**, 661–668 (2004).

125. M. Zacharias, Accounting for conformational changes during protein-protein docking. *Curr. Opin. Struct. Biol.* **20**, 180–186 (2010).

126. S. E. Dobbins, V. I. Lesk, M. J. E. Sternberg, Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc. Natl. Acad. Sci. USA* **105**, 10390–10395 (2008).

127. A. Stein, M. Rueda, A. Panjkovich, M. Orozco, P. Aloy, A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. *Structure* **19**, 881–889 (2011).

128. R. Chaudhuri, O. Carrillo, C. A. Laughton, M. Orozco, Application of drug-perturbed essential dynamics/molecular dynamics (ED/MD) to virtual screening and rational drug design. *Journal of Chemical Theory and Computation* **8**, 2204–2214 (2012).

129. L. Orellana, A. Hospital, M. Orozco, *Oncogenic mutations of the EGF-Receptor ectodomain reveal an unexpected mechanism for ligand-independent activation* (biorxiv, 2014).

130. M. K. Kim, G. S. Chirikjian, R. L. Jernigan, Elastic models of conformational transitions in macromolecules. *J. Mol. Graphics Modell.* **21**, 151–160 (2002).

131. M. K. Kim, R. L. Jernigan, G. S. Chirikjian, Efficient generation of feasible pathways for protein conformational transitions. *Biophys. J.* **83**, 1620–1630 (2002).

132. L. Yang, G. Song, R. L. Jernigan, How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys. J.* **93**, 920–929 (2007).

133. D. Tobi, I. Bahar, Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. USA* **102**, 18908–18913 (2005).

134. A. Bakan, I. Bahar, The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. USA* **106**, 14349–14354 (2009).

135. P. Doruker, R. L. Jernigan, I. Bahar, Dynamics of large proteins through hierarchical levels of coarse-grained structures. *Journal of Computational Chemistry* **23**, 119–127 (2002).

136. I. Bahar, C. Chennubhotla, D. Tobi, Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr. Opin. Struct. Biol.* **17**, 633–640 (2007).

137. Z. Yang, P. Májek, I. Bahar, Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PLoS Comput Biol* **5**, e1000360 (2009).

138. R. Mendez, U. Bastolla, Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. *Phys. Rev. Lett.* **104**, 228103 (2010).

139. I. Bahar, R. L. Jernigan, Inter-residue Potentials in Globular Proteins and the Dominance of Highly Specific Hydrophilic Interactions at Close Separation. *J. Mol. Biol.* **266**, 195–214 (1997).

140. A. V. Sinitskiy, G. A. Voth, Coarse-graining of proteins based on elastic network models. *Chem. Phys.* **422**, 165–174 (2013).

141. V. Tozzini, Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **15**, 144–150 (2005).

142. C. Clementi, Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **18**, 10–15 (2008).

143. M. G. Saunders, G. A. Voth, Coarse-graining of multiprotein assemblies. *Curr. Opin. Struct. Biol.* **22**, 144–150 (2012).

144. I. Buch, T. Giorgino, G. De Fabritiis, Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **108**, 10184–10189 (2011).

145. N. Stanley, S. E.-M. I. n, G. De Fabritiis, Kinetic modulation of a disordered protein domain by phosphorylation. *Nature*

Communications **5**, 1–8 (2014).

146. D. T. Major, J. Gao, A combined quantum mechanical and molecular mechanical study of the reaction mechanism and alpha-amino acidity in alanine racemase. *J. Am. Chem. Soc.* **128**, 16345–16357 (2006).

147. R. Molina *et al.*, Visualizing phosphodiester-bond hydrolysis by an endonuclease. *Nat Struct Mol Biol* **22**, 65–72 (2015).

148. P. C. Loewen, X. Carpena, P. Vidossich, I. Fita, C. Rovira, An ionizable active-site tryptophan imparts catalase activity to a peroxidase core. *J. Am. Chem. Soc.* **136**, 7249–7252 (2014).

149. C. Hensen *et al.*, A combined QM/MM approach to protein--ligand interactions: polarization effects of the HIV-1 protease on selected high affinity inhibitors. *Journal of medicinal chemistry* **47**, 6673–6680 (2004).

150. K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, D. E. Shaw, Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* **134**, 3787–3791 (2012).

151. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How Fast-Folding Proteins Fold. *Science* **334**, 517–520 (2011).

152. D. E. Shaw *et al.*, Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **330**, 341–346 (2010).

153. J. Vreede, J. Juraszek, P. G. Bolhuis, Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein. *Proc. Natl. Acad. Sci. USA* **107**, 2397–2402 (2010).

154. L. S. Stelzl, P. W. Fowler, M. S. P. Sansom, O. Beckstein, Flexible Gates Generate Occluded Intermediates in the Transport Cycle of LacY. *J. Mol. Biol.* **426**, 735–751 (2014).

155. T. Shimamura *et al.*, Molecular Basis of Alternating Access Membrane Transport by the Sodium-Hydantoin Transporter Mhp1. *Science* **328**, 470–473 (2010).

156. D. Röthlisberger *et al.*, Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).

157. G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker, K. N. Houk, Computational Enzyme Design. *Angew. Chem. Int. Ed.* **52**, 5700–5725 (2013).

158. A. D. Pearson *et al.*, Trapping a transition state in a computationally designed protein bottle. *Science* **347**, 863–867 (2015).

159. D. P. Tieleman *et al.*, Membrane protein simulations with a united-atom lipid and all-atom protein model: lipid-protein interactions, side chain transfer free energies and model proteins. *J Phys Condens Matter* **18**, S1221–34 (2006).

160. P. Sfriso *et al.*, Finding Conformational Transition Pathways from Discrete Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* **8**, 4707–4718 (2012).

161. A. Fernández, A. Colubri, Pathway heterogeneity in protein folding. *Proteins: Struct., Funct., Bionf.* **48**, 293–310 (2002).

162. G. Wei, N. Mousseau, P. Derreumaux, Complex folding pathways in a simple beta-hairpin. *Proteins: Struct., Funct., Bionf.* **56**, 464–474 (2004).

163. Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, P. G. Wolynes, Optimizing physical energy functions for protein folding. *Proteins: Struct., Funct., Bionf.* **54**, 88–103 (2004).

164. T. Vuorela *et al.*, Role of lipids in spheroidal high density lipoproteins. *PLoS Comput Biol* **6**, e1000964 (2010).

165. Y. Oguchi *et al.*, A tightly regulated molecular toggle controls AAA+ disaggregase. *Nat Struct Mol Biol* **19**, 1338–1346 (2012).

166. S. Yang *et al.*, Domain swapping is a consequence of minimal frustration. *Proc. Natl. Acad. Sci. USA* **101**, 13786–13791 (2004).

167. Y. Levy, P. G. Wolynes, J. N. Onuchic, Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. USA* **101**, 511–516 (2004).

168. M. Levitt, A. Warshel, Computer simulation of protein folding. *Nature* **253**, 694–698 (1975).

169. S. Brown, N. J. Fawzi, T. Head-Gordon, Coarse-grained sequences for protein folding and design. *Proc. Natl. Acad. Sci. USA* **100**, 10712–10717 (2003).

170. Y. C. Kim, G. Hummer, Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. *J. Mol. Biol.* **375**, 1416–1433 (2008).

171. G. S. Ayton, G. A. Voth, Multiscale computer simulation of the immature HIV-1 virion. *Biophys. J.* **99**, 2757–2765 (2010).

172. D. Flöck, V. Helms, A Brownian dynamics study: the effect of a membrane environment on an electron transfer system. *Biophys. J.* **87**, 65–74 (2004).

173. A. H. Elcock, Atomic-level observation of macromolecular crowding effects: escape of a protein from the GroEL cage. *Proc. Natl. Acad. Sci. USA* **100**, 2340–2344 (2003).

174. S. R. McGuffee, A. H. Elcock, J. M. Briggs, Ed. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput Biol* **6**, e1000694 (2010).

175. J. Balbo, P. Mereghetti, D.-P. Herten, R. C. Wade, The shape of protein crowders is a major determinant of protein diffusion. *Biophys. J.* **104**, 1576–1584 (2013).

176. M. W. Mahoney, W. L. Jorgensen, A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **112**, 8910 (2000).

177. S. W. Rick, A reoptimization of the five-site water potential (TIP5P) for use with Ewald sums. *J. Chem. Phys.* **120**, 6085 (2004).

178. W. L. Jorgensen, J. Chandrasekhar, J. Madura, Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **79**, 926–938 (1983).

179. J. L. F. Abascal, C. Vega, A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **123**, 234505 (2005).

180. J. L. F. Abascal, E. Sanz, R. García Fernández, C. Vega, A potential model for the study of ices and amorphous water: TIP4P/Ice. *J. Chem. Phys.* **122**, 234511 (2005).

181. H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, The missing term in effective pair potentials. *J. Phys. Chem.* … **91**, 6269–6271 (1987).

182. M. Orozco, F. J. Luque, Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.* **101**, 203–204 (2001).

183. J. Chen, C. L. Brooks, J. Khandogin, Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.* **18**, 140–148 (2008).

184. N. A. Baker, Improving implicit solvent simulations: a Poisson-centric view. *Curr. Opin. Struct. Biol.* **15**, 137–143 (2005).

185. J. Kleinjung, F. Fraternali, Design and application of implicit solvent models in biomolecular simulations. *Curr. Opin. Struct. Biol.* **25**, 126–134 (2014).

186. B. Zagrovic, E. J. Sorin, V. Pande, β-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.* **313**, 151–169 (2001).

187. L. Darré, M. R. Machado, S. Pantano, Coarse-grained models of water. *WIREs Comput. Mol. Sci.* **2**, 921–930 (2012).

188. S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta, S. J. Marrink, Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comput Biol* **6**, e1000810 (2010).

189. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. De Vries, The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* **111**, 7812–7824 (2007).

190. J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, M. L. Klein, A Coarse Grain Model for Phospholipid Simulations. *J. Phys. Chem. B* **105**, 4464–4470 (2001).

191. J. C. Shelley *et al.*, Simulations of Phospholipids Using a Coarse Grain Model. *J. Phys. Chem. B* **105**, 9785–9792 (2001).

192. Z.-J. Wang, M. Deserno, A Systematically Coarse-Grained Solvent-Free Model for Quantitative Phospholipid Bilayer Simulations. *J. Phys. Chem. B* **114**, 11207–11220 (2010).

193. Z. Wu, Q. Cui, A. Yethiraj, A New Coarse-Grained Model for Water: The Importance of Electrostatic Interactions. *J. Phys. Chem. B* **114**, 10524–10529 (2010).

194. S. Riniker, W. F. van Gunsteren, A simple, efficient polarizable coarse-grained water model for molecular dynamics simulations. *J. Chem. Phys.* **134**, 084110 (2011).

195. M. Orsi, J. W. Essex, The ELBA Force Field for Coarse-Grain Modeling of Lipid Membranes. *PloS one* **6**, e28637 (2011).

196. L. Darré, M. R. Machado, P. D. Dans, F. E. Herrera, S. Pantano, Another Coarse Grain Model for Aqueous Solvation: WAT FOUR? *Journal of Chemical Theory and Computation* **6**, 3793–3807 (2010).

197. C. Zhang, J. Ma, Enhanced sampling and applications in protein folding in explicit solvent. *J. Chem. Phys.* **132**, 244101–244101–16 (2010).

198. V. Spiwok, Z. Šućur, P. Hošek, Biotechnology Advances. *Biotechnology Advances* **33**, 1130–1140 (2014).

199. J. Schlitter, M. Engels, P. Krüger, Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *Journal of molecular graphics* **12**, 84–89 (1994).

200. S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, K. Schulten, Molecular Dynamics Study of Unbinding of the Avidin-Biotin Complex. *Biophys. J.* **72**, 1568–1581 (1997).

201. B. Isralewitz, M. Gao, K. Schulten, Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.*, 224–230 (2001).

202. H. Grubmüller, B. Heymann, P. Tavan, Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **271**, 997–999 (1996).

203. C. Jarzynski, Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **78**, 2690–2693 (1997).

204. A. F. Voter, Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **78**, 3908 (1997).

205. D. Hamelberg, J. Mongan, J. A. McCammon, Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919 (2004).

206. Y. Miao *et al.*, Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *Journal of Chemical Theory and Computation* **10**, 2677–2689 (2014).

207. Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**, 141–151 (1999).

208. D. J. Earl, M. W. Deem, Parallel tempering: Theory, applications, and new perspectives. *Phys Chem Chem Phys* **7**, 3910 (2005).

209. S. Kumar, J. Rosenberg, D. Bouzida, R. Swendsen, P. Kollman, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **13**, 1011–1021 (1992).

210. G. M. Torrie, J. P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).

211. F. Zhu, G. Hummer, Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry* **33**, 453–465 (2011).

212. S. Krumar, D. Bouzida, R. H. Swendsen, P. Kollman, J. Rosenberg, The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. *Journal of Computational Chemistry* **13**, 1011–1021 (1992).

213. P. G. Bolhuis, Transition-path sampling of beta-hairpin folding. *Proc. Natl. Acad. Sci. USA* **100**, 12129–12134 (2003).

214. C. Dellago, P. Bolhuis, P. L. Geissler, Transition path sampling. *Adv. Chem. Phys.* **123**, 1–78 (2002).

215. A. K. Faradjian, R. Elber, Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **120**, 10880 (2004).

216. R. Elber, A Milestoning Study of the Kinetics of an Allosteric Transition: Atomically Detailed Simulations of Deoxy Scapharca Hemoglobin. *Biophys. J.* **92**, L85–L87 (2007).

217. R. Elber, A. West, Atomically detailed simulation of the recovery stroke in myosin by Milestoning. *Proc. Natl. Acad. Sci. USA* **107**, 5001–5005 (2010).

218. R. Elber, Long-timescale simulation methods. *Curr. Opin. Struct. Biol.* **15**, 151–156 (2005).

219. P. Májek, R. Elber, Milestoning without a Reaction Coordinate. *Journal of Chemical Theory and Computation* **6**, 1805–1817 (2010).

220. M. Rueda, E. Cubero, C. A. Laughton, M. Orozco, Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations? *Biophys. J.* **87**, 800–811 (2004).

221. D. M. Zuckerman, T. B. Woolf, Dynamic reaction paths and rates through importance-sampled stochastic dynamics. *J. Chem. Phys.* **111**, 9475 (1999).

222. Supplementary Material: Zipping and Unzipping of Adenylate Kinase: Atomistic Insigths inyo the Ensemble of Open Closed Transitions. 1–19 (2009).

223. A. Laio, M. Parrinello, Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99**, 12562–12566 (2002).

224. A. Barducci, M. Bonomi, M. Parrinello, Metadynamics. *WIREs Comput. Mol. Sci.* **1**, 826–843 (2011).

225. V. Leone, F. Marinelli, P. Carloni, M. Parrinello, Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.* **20**, 148–154 (2010).

226. A. Barducci, G. Bussi, M. Parrinello, Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100**, 020603 (2008).

227. S. Vega, O. Abian, A. Velazquez-Campoy, On the link between conformational changes, ligand binding and heat capacity. *Biochim. Biophys. Acta*, 1–11 (2015).

228. P. Csermely, R. Palotai, R. Nussinov, Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in Biochemical Sciences* **35**, 539–546 (2010).

229. S. Hammes-Schiffer, S. J. Benkovic, Relating protein motion to catalysis. *Annu. Rev. Biochem.* **75**, 519–541 (2006).

230. Y. Shan, A. Arkhipov, E. T. Kim, A. C. Pan, D. E. Shaw, Transitions to catalytically inactive conformations in EGFR kinase. *Proc. Natl. Acad. Sci. USA* **110**, 7270–7275 (2013).

231. N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. I. Shakhnovich, Discrete molecular dynamics studies of the folding of a protein-like model. *Folding Des.* **3**, 577–587 (1998).

232. M. B. Kubitzki, B. L. de Groot, The atomistic mechanism of conformational transition in adenylate kinase: a TEE-REX molecular dynamics study. *Structure* **16**, 1175–1182 (2008).

233. D. Seeliger, B. L. de Groot, T. Lengauer, Ed. Conformational Transitions upon Ligand Binding: Holo-Structure Prediction from Apo Conformations. *PLoS Comput Biol* **6**, e1000634 (2010).

234. B. Różycki, Y. C. Kim, G. Hummer, SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* **19**, 109–116 (2011).

235. P.-C. Chen, J. S. Hub, Validating Solution Ensembles from Molecular Dynamics Simulation by Wide-Angle X-ray Scattering Data. *Biophys. J.* **107**, 435–447 (2014).

236. F. Morcos, B. Jana, T. Hwa, J. N. Onuchic, Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. USA* **110**, 20533–20538 (2013).

237. Y. Ye, A. Godzik, FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nuc. Acids Res.* **32**, W582–W585 (2004).

238. D. R. Weiss, M. Levitt, Can morphing methods predict intermediate structures? *J. Mol. Biol.* **385**, 665–674 (2009).

239. N. Echols, D. Milburn, M. Gerstein, MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nuc. Acids Res.* **31**, 478–482 (2003).

240. E. Lindahl, C. Azuara, P. Koehl, M. Delarue, NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nuc. Acids Res.* **34**, W52–W56 (2006).

241. A. Ahmed, F. Rippmann, G. Barnickel, H. Gohlke, A Normal Mode-Based Geometric Simulation Approach for Exploring Biologically Relevant Conformational Transitions in Proteins. *J. Chem. Inf. Model* **51**, 1604–1622 (2011).

242. D. Seeliger, J. Haas, B. L. de Groot, Geometry-based sampling of conformational transitions in proteins. *Structure* **15**, 1482–1492 (2007).

243. C. Hyeon, J. N. Onuchic, Internal strain regulates the nucleotide binding site of the kinesin leading head. *Proc. Natl. Acad. Sci. USA* **104**, 2175–2180 (2007).

244. D. S. Marks, T. A. Hopf, C. Sander, Protein structure prediction from sequence variation. *Nature Biotechnology* **30**, 1072–1080 (2012).

245. D. S. Marks *et al.*, A. Sali, Ed. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PloS one* **6**, e28766

(2011).

246. T. A. Hopf *et al.*, Three-Dimensional Structures of Membrane Proteinsfrom Genomic Sequencing. *Cell* **149**, 1607–1621 (2012).

247. T. A. Hopf *et al.*, Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014).

248. D. de Juan, F. Pazos, A. Valencia, Emerging methods in protein co-evolution. *Nat Rev Genet* **14**, 249–261 (2013).

249. A. Bakan *et al.*, Evol and ProDy for bridging protein sequence evolution and structural dynamics. *Bioinformatics* **30**, 2681–2683 (2014).

250. J. L. MacCallum, A. Pérez, K. A. Dill, Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. USA* **112**, 6985–6990 (2015).

251. A. Hospital *et al.*, MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics* **28**, 1278–1279 (2012).

252. R. J. Allen, P. B. Warren, P. R. ten Wolde, Sampling Rare Switching Events in Biochemical Networks. *Phys. Rev. Lett.* **94**, 018104 EP– (2005).

253. P. Sfriso, A. Hospital, A. Emperador, M. Orozco, Exploration of conformational transition pathways from coarse-grained simulations. *Bioinformatics* **29**, 1980–1986 (2013).

254. T. T. Foley, M. S. Shell, W. G. Noid, The impact of resolution upon entropy and information in coarse-grained models. *J. Chem. Phys.* **143**, 243104 (2015).

255. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. USA* **110**, 15674–15679 (2013).

256. A. C. Pan, T. M. Weinreich, Y. Shan, D. P. Scarpazza, D. E. Shaw, Assessing the Accuracy of Two Enhanced Sampling Methods Using EGFR Kinase Transition Pathways: The Influence of Collective Variable Choice. *Journal of Chemical Theory and Computation* **10**, 2860–2865 (2014).